Positive ID

Towards better cancer cell–line authentication and experimental reproducibility

PAGES 264 & 307
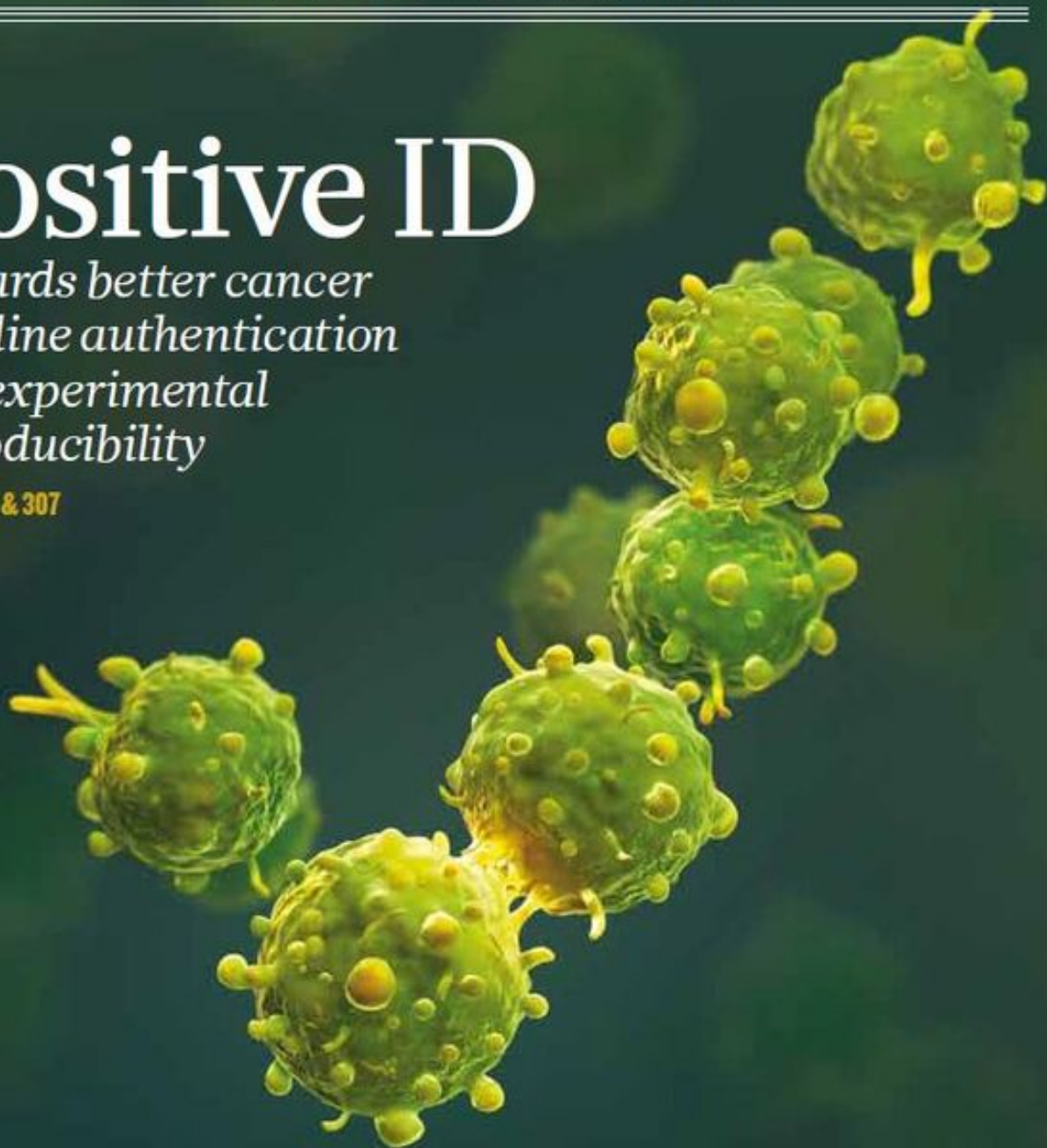
HUBBLE AT 25

STAR PERFORMER

The space telescope is a tough act to follow

PAGES 282 & 287

CANCER THERAPY

TARGET SELECTION

Druggable Ras and precision care in sight

PAGES 278 & 290

PLANETARY SCIENCE

A TOUCH OF MERCURY

The missing ingredient in Earth's mantle

PAGES 299 & 337

THIS WEEK

Close the deal

Science–based evidence and enlightened diplomacy have brought within reach a historic opportunity for nuclear détente with Iran. It must be seized.

In a diplomatic tour de force, negotiators from six world powers and Iran reached — against the odds — a tentative agreement on 2 April to ensure that Iran's nuclear programme is for peaceful purposes only. It is science-informed diplomacy and foreign policy at its best. Even the most optimistic of seasoned nuclear-weapons and non-proliferation experts were surprised by the comprehensiveness of the interim accord, its level of detail, and the substantial concessions made on both sides. Few had expected this degree of progress given the decades of hostility and intransigence on both sides.

In a perhaps unprecedented flurry of published opinion pieces and statements, experts have overwhelmingly lent their support to the accord. They have also subjected it to robust online peer review, highlighting the positive outcomes, but also pointing out the technical loopholes and details that they feel must still be negotiated or clarified before the 30 June deadline for a final agreement.

The emphasis on getting the scientific and technical assessments that underpin the issues right, to offer political leaders confidence in the projected outcomes, has played a central part in getting to this crucial juncture. Two physicists, both at the Massachusetts Institute of Technology in Cambridge in the 1970s, had key roles: Ali Akbar Salehi, head of Iran's Atomic Energy Organization, and Ernest Moniz, the US energy secretary. In long face-to-face discussions, the men thrashed out the complex nuclear science to come up with acceptable compromises that did not cross the red lines of either side. Importantly, the lead negotiators, US secretary of state John Kerry and Iranian foreign minister Mohammad Javad Zarif, have taken on board their scientific advice.

Scientists share a language, culture and values that can help to transcend politics and enmity. Researchers involved in past nuclear-weapons treaties say that scientific collaboration between adversaries is crucial to building trust and confidence, but they emphasize that it takes time. Iran has been ostracized by many governments for almost four decades, and rebuilding trust on both sides will take years. That this rapprochement is now under way can only be commended — especially at this time of exceptional political instability in the Middle East, which has unexpectedly aligned some of Iran's and the West's strategic interests. Any easing of the sanctions on Iran and its political isolation will also benefit the country's isolated scientific community.

Experts are unanimous that the framework of the deal shows that it could essentially put Iran's nuclear programme on ice for well over a decade — and so buy the time needed to build greater trust and to develop further measures to ensure that any eventual larger Iranian nuclear programme remains peaceful. The accord would, for example, block Iran's potential route to a plutonium bomb, by redesigning the country's Arak heavy-water reactor to make it much less capable of producing weapons-grade plutonium. Moreover, all plutonium-containing spent fuel would be shipped out of the country.

Iran's potential to make a bomb using enriched uranium would also be curtailed to the extent judged necessary by scientists to ensure that, for the foreseeable future, it would take the country more than a year to 'break out' and develop a nuclear weapon, leaving enough time for international intervention (see page 274). The Vienna-based International Atomic Energy Agency would also be given unprecedented powers to inspect Iran's entire nuclear programme for 25 years to ensure that, were Iran to violate the agreement either overtly or covertly, this would be detected quickly.

> *"Scientists share a language, culture and values that can help to transcend politics."*

Iranian and other scientists emphasize the interplay of science and politics. The breakthrough was made possible, they say, only by the election of the relative reformer Hassan Rouhani to the Iranian presidency in 2013 and of Barack Obama to the US presidency in 2008. Both leaders have been more open to pragmatic and constructive dialogue between the two nations than their predecessors. Critics of the deal have yet to put forward any credible alternatives, or any substantive challenges to its technical underpinnings, relying rather on political rhetoric and stoking fear to justify inaction. The late US president Ronald Reagan famously adopted the Russian proverb "Trust, but verify" with respect to the monitoring of nuclear-disarmament treaties with the Soviet Union. It is time once again for progressive policies to prevail over dangerous inaction. ∎

Numbers matter

Researchers need help in making the statistical power of animal experiments clear.

Albert Einstein is said to have noted that theories should be as simple as possible, but no simpler. By the same token, biomedical researchers doing *in vivo* experiments should use as few animals as possible, but no fewer. On page 271, *Nature* reports a move by UK government funding agencies to require grant applicants to show how they calculated the number of animals needed to make the results of an experiment statistically robust. In recent years there have been concerns that sample sizes in individual experiments can be too low, especially in preclinical research that attempts to determine whether a drug is worth pursuing in human studies.

Too-small sample sizes can lead to promising drugs being discarded when their effectiveness is missed, or to false positives, as well as to ethical issues if animals are being used in studies that are too small to provide reliable results.

The UK research councils' move is to be applauded. And Britain is

not alone in pursuing such improvements: the US National Institutes of Health has been testing the use of a grant-review checklist that includes features such as experimental design, to improve the reproducibility of preclinical research in animals.

The burden for this should not fall on funding bodies alone. Institutions must also increase the amount of support offered to researchers in designing the statistical aspects of an experiment. Such support is too often limited or ad hoc: study design is complex and needs careful consideration by people who truly understand the issues (see *Nature* **506,** 131–132; 2014).

Journals are also responsible for ensuring that the research they publish is reported in sufficient detail for readers to fully appreciate key details of experimental and analytical design. Many publications — including *Nature* — have endorsed the ARRIVE guidelines for reporting animal research (C. Kilkenny *et al. PLoS Biol.* **8,** e1000412; 2010). These are, however, hugely detailed, and compliance at this level is difficult for early, exploratory research.

Journals published by Nature Publishing Group nevertheless encourage the use of ARRIVE. In 2013, we implemented a reporting checklist that demands that authors supply key details of study design. For animal studies, these include the methods of sample-size determination, randomization and study blinding, as well as exclusion criteria (see *Nature* **496,** 398; 2013). An impact analysis on the effectiveness of the changes introduced in 2013 is currently under way.

Sample size is just one of a suite of issues that need to be addressed if poor reproducibility is to be tackled. Journals have a key part to play in dealing with this problem, but so do others. Credit to

those academies that take a lead. This month, for example, the UK Academy of Medical Sciences held a meeting in London at which researchers, funders and representatives from research institutions and universities attempted to provide recommendations for improving reproducibility by examining case studies in disciplines from epidemiology to particle physics, and by exploring the role of culture

"There are no magic bullets — all parts of the research community need to chip away at the problem."

and incentives. There are no magic bullets — all parts of the research community need to chip away at the problem.

Undoubtedly, part of the challenge is the culture that pushes investigators in many parts of the world to produce more and more with the same resources. The drive to maximize the number of papers and the impact of findings is pervasive.

In a commentary published in *Nature Biotechnology* last year, experimental psychologist Marcus Munafò and his colleagues compared modern biomedical research with the 1970s automobile industry (M. Munafò *et al. Nature Biotechnol.* **32,** 871–873; 2014). The fast-moving but error-prone car production lines of the United States found themselves losing ground to Japanese manufacturers that stressed the importance of quality-control at every step in their factories.

The moral of the story: quality assurance adds a burden, but it is worth the effort for a longer-term gain in public confidence. Making sure that the power of an animal experiment suits its purpose is an important way for funders and researchers to contribute. ∎

Time to tackle cells' mistaken identity

The differences between a cow and a monkey are clear. It is easy to tell a moth from a mosquito. So why are there still scientific studies that mix them up? The answer is simple: hundreds of cell lines stored and used by modern laboratories have been wrongly identified. Some pig cells are labelled as coming from a chicken; cell lines advertised as human have been shown to contain material from hamsters, rats, mice and monkeys.

Which is worse: that such crude mix-ups exist, or that, every day, researchers use cell lines that somebody, somewhere has already found to be mislabelled, misidentified or contaminated? To solve the first problem is a huge challenge. To address the second is a more manageable task, and one that researchers, journals, universities and funders must take seriously.

Nature and the Nature research journals are strengthening their policies to improve the situation. From next month, we will ask authors to check that they are not working on cells known to have been misidentified or cross-contaminated, and will ask them to provide more details about the source and testing of their cell lines.

This may sound like an obvious way to deal with a problem that has been known about for decades. But tests to check the contents of cell lines are complex and time-consuming, and until recently were expensive. What makes the time ripe for action is a combination of a rising awareness of the problem among scientists in certain communities (cancer research in particular), the availability of proper tests and resources (see J. R. Masters *Nature* **492,** 186 (2012), and page 307), and the willingness of some funders to tackle the matter — including the US National Institutes of Health and the Prostate Cancer Foundation in Santa Monica, California.

Problems have already been found with more than 400 cell lines. In the long term, the goal must be to change testing routines worldwide to ensure that new mix-ups are not propagated. The least that scientists should already be doing is checking whether the cell line they are using is one of those already marked with a red flag.

In 2013, Nature journals started to ask authors to report the source of their cell line and whether the cell line had been authenticated. Most have not done so. Out of a sample of around 60 cell-line-based papers published across several Nature journals in the past two years, almost one-quarter did not report the source. Only 10% of authors said that they had authenticated the cell line. This is especially problematic given that almost one-third said that they had obtained the cell lines as a gift from another laboratory.

From 1 May, all authors of papers involving cell lines that are submitted to Nature journals will be asked whether they have checked their cell lines against publicly available lists of those known to be problematic. We will in particular monitor compliance in cancer research. The focus on cancer is a first step, chosen because the cell-line problem has been best documented in this field, and because the cancer community is already reacting to the issue. Some specialist journals, such as the *International Journal of Cancer*, are now systematically asking for authentication. This is important not only for its effects on basic research, but also because of the potential for translational research to founder if cell lines are contaminated.

Other fields are not immune to cell-line problems, and we hope to extend the systematic checks to them in future. More details of the new policy, whom it affects and where the cell lines should be checked are available at go.nature.com/zqjubh.

That a cell line used in a research project appears on a watch-list need not make the research invalid, or mean that the paper will automatically be rejected. Authors will be asked to explain why the misidentification does not undermine the conclusions. But we reserve the right to ask for data to be removed if the justification is judged insufficient by editors and referees. ∎

GUARDIAN NEWS & MEDIA

Scientists must speak up on fossil–fuel divestment

Alan Rusbridger *wants researchers to help convince powerful philanthropic organizations to set an example and stop propelling carbon emissions.*

Editors call it 'eat-your-peas' journalism — stories that are really good for you, if not nearly as enjoyable as the latest news about Jeremy Clarkson or the wardrobe malfunction of a breakfast television presenter.

Climate change is the ultimate eat-your-peas journalism. On some level, most people are aware that they should be deeply concerned about it. On another level, they just aren't. Perhaps it is just too frightening to think about. The story changes little from day to day. And, anyway, there seems to be little that anyone can do about it. A depressing fatalism settles over the subject. News editors shrug and change the subject.

But what if the climate story is the most important news on Earth — in the sense that, if we can't find a solution, then our children and grandchildren may well inherit a planet that is deeply hostile to the sort of civilization we enjoy?

I pondered this question at home over Christmas. I had been editing *The Guardian* for nearly 20 years and had announced that I would step down in the summer of 2015. Was there — in my time still left as editor — the opportunity to do something sharp and focused about climate change? Something that would make people wolf down their peas with relish?

I had in my mind the words of the US writer and environmental campaigner Bill McKibben: this thing has moved beyond the environment pages. The scientists and ecologists have done brilliant work over the years, but the essentials are now settled. The climate story has moved into the realms of politics, finance and economics. That is how you would have to write the story to make an impact.

Newspaper campaigns can energize and inspire people in a way that simple reporting sometimes does not. *The Guardian* toyed with the idea of aiming such a campaign at policy-makers, but that felt more like eating broccoli. It would have been easy, but probably not effective, to aim at the big, bad and familiar targets in the fossil-fuel industries.

McKibben convinced us to focus on the three numbers that could determine the future of our species. The first, 2 °C, is the internationally agreed warming threshold for dangerous climate-change impacts. The second figure is the amount of extra carbon dioxide emissions that are likely to push us over that threshold. The final figure is the amount of carbon dioxide that would be produced if all of the known fossil-fuel reserves in the world were extracted and burned.

There is, of course, uncertainty around these numbers. And as we burn fossil fuels ever faster they present a moving target. But what is very clear is that the third figure is much higher than the second — three to five times higher, in fact. Therefore the majority of the oil, gas and coal reserves can never be allowed to be dug up. And fossil-fuel companies should not waste investor capital prospecting for more such reserves.

Companies with these reserves are almost certainly vastly overvalued, and this is dawning on a great many people — from central bankers to investment-fund managers, faith leaders, chief executives, universities and non-governmental organizations.

But not everyone agrees on how to respond. Some protest that divesting from fossil fuels will simply lead to 'bad' money replacing 'good'. Or that they have a duty to maximize returns. Or that keeping money in these companies enables 'good' people to 'engage' and have some influence.

Somewhat surprisingly, there are some 'good' organizations that have so far declined to move their money out of oil, gas and coal. There are few better foundations in the fields of science and medicine than the Bill & Melinda Gates Foundation and the Wellcome Trust. They give away huge amounts of money to projects and research that save countless lives and advance human knowledge and understanding. There is almost nothing not to like about them.

> SOME PROTEST THAT DIVESTING FROM **FOSSIL FUELS** WILL SIMPLY LEAD TO **'BAD' MONEY** REPLACING **'GOOD'.**

But neither foundation will take their money out of the companies that cannot be allowed to extract and burn all the hydrocarbons they own.

And so, as part of our campaign, Keep it in the Ground, we have asked these organizations — politely and respectfully, but with determination — to think again. More than 180,000 readers have signed a petition asking them to reconsider. And, if you were about to ask, the Guardian Media Group has, in the space of two months, moved from not really thinking very much about the issue to announcing that its £800-million (US$1.2-billion) fund will divest from fossil fuels within 2–5 years.

Wellcome's excuse — that it prefers to "engage" with the fossil-fuel giants — sounds feeble. It has not produced any evidence of tangible gains from the strategy. If Wellcome can genuinely point to the fruits of engagement, it should surely — like good scientists — demonstrate the evidence, not hide behind commercial confidentiality.

Likewise, if the Gates wants to demonstrate that the good it does outweighs the harmful activities it helps to fund, it should come out and make that case public.

In the absence of such evidence, these wonderful progressive foundations are failing to show the kind of leadership that could be transformative in shifting policy arguments and influencing others. The voices that will resonate loudest with the Wellcome and the Gates are those of scientists. I urge you to make them heard. ∎

Alan Rusbridger *is editor-in-chief of* The Guardian *in London.*
e-mail: alan.rusbridger@theguardian.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

Neutrinos from a galaxy far away

Two of the most energetic neutrinos detected by a telescope in the Antarctic may have come from the cores of distant galaxies.

Neutrinos are stable and can travel far in space, so they could shed light on distant astrophysical and galactic objects. The Antarctic telescope IceCube picked up signs of neutrinos in 2011 and 2012 that were the first ever measured with energies of 1 petaelectronvolt (1×10^{15} electronvolts), suggesting a powerful source such as a blazar — a type of high-energy galaxy.

A team led by Clancy James of the University of Erlangen and Matthias Kadler of the University of Würzburg, both in Germany, studied six years of data from the underwater ANTARES neutrino telescope off the coast of Toulon, France, scanning six blazars for further neutrinos. The two blazars considered to be the best candidates each yielded events that were consistent with the signature of a neutrino, suggesting that they could be the sources of the IceCube neutrinos.

Astron. Astrophys. **576**, L8 (2015)

Brain zap stops electrical fault

Deep-brain stimulation may improve movement in people with Parkinson's disease by reducing abnormally strong coupling of electrical activity in the brain.

Implanted electrodes are used to treat some brain disorders, particularly Parkinson's disease. Coralie de Hemptinne at the University of California, San Francisco, and her colleagues recorded electrical potentials in the motor cortex of 23 people with Parkinson's who were undergoing surgery to implant electrodes into their brains. The researchers found that when they switched the electrodes on, the coupling of electrical activity in the motor cortex was reduced, and that the level of uncoupling correlated with the degree to which the patients' movements improved.

The authors say that the results could inform the design of improved devices for deep-brain stimulation.

Nature Neurosci. http://dx.doi.org/10.1038/nn.3997 (2015)

A 3D map of skin microbes and molecules

Researchers have glimpsed the complexities of human skin by creating a three-dimensional (3D) map of the chemicals and microbes found on the body's largest organ.

Pieter Dorrestein of the University of California in San Diego and his colleagues swabbed 400 locations on the skin of two healthy human volunteers who abstained from bathing for three days before sampling. Using mass spectrometry and DNA sequencing, the researchers identified the chemical compounds and microbes on the skin. They used a supercomputer to combine the data and to build a map covering the whole body (pictured is the chemical map for one volunteer; blue is low molecular diversity, red is high).

The team now plans to characterize more skin chemicals and microbes, and say that their technique could be used in fields from forensics to beauty-product development.

Proc. Natl Acad. Sci. USA http://doi.org/3h8 (2015)

THEODORE ALEXANDROV/*PROC. NATL. ACAD. SCI. USA*

Fishing drives population decline

Fishing magnifies natural variations in numbers of fish, increasing the risk of population collapses.

Timothy Essington and his colleagues at the University of Washington in Seattle analysed at least 25 years' worth of data on 55 stocks of small fish such as sardines, herrings and anchovies that are preyed on by others. The population sizes of these species fluctuate naturally and widely over time. But the researchers found that when populations collapsed to less than 25% of their mean size, the stocks were more likely to have experienced exceptionally high fishing rates before the collapse than to have seen large natural variations in size.

Modelling the fish populations suggests that fishery management practices that do not respond quickly to dips in species numbers

increase both the magnitude and frequency of natural population declines.
Proc. Natl Acad. Sci. USA http://doi.org/3hk (2015)

SEISMOLOGY

San Francisco's quake hazard rises

Two geological faults in northern California are linked, meaning that the risk of a large earthquake in the eastern San Francisco Bay Area is greater than was thought.

A team led by Estelle Chaussard of the University of California, Berkeley, used satellite radar to study ground deformation along the Hayward fault, east of San Francisco. The scientists found that it connected with the Calaveras fault. Both are part of the San Andreas system and were considered to be separate.

The combined fault system could unleash an earthquake greater than magnitude 7, bigger than had been expected.
Geophys. Res. Lett. http://doi.org/3hh (2015)

ATMOSPHERIC SCIENCE

Asian pollution hitchhikes south

Pollution from East Asia affects air quality in the distant tropics.

A team led by Matthew Ashfold at the University of Cambridge, UK, detected elevated levels of a chlorine-containing gas at two remote sites in tropical Borneo during the Northern Hemisphere winter of 2008–09. The team used an atmospheric transport model to show that the chemical — an indicator of a range of industrial pollutants — was transported southward

NASA

PETAR KUJUNDZIC/REUTERS/CORBIS

from east Asia by rapidly moving cold air masses.

During cold surges, east Asian air pollution (**pictured**) can reach the equator in a few days. If ozone-degrading chlorine pollutants are lifted by convection into the tropical atmosphere, even short-lived compounds might have a negative effect on stratospheric ozone, the authors say.
Atmos. Chem. Phys. **15**, 3565–3573 (2015)

MICROBIOLOGY

Downsides of low-dose antibiotics

Taking low doses of antibiotics to prevent recurring bladder infections could make the illness worse than taking no antibiotic at all.

Lee Goneau of the University of Toronto in Canada and his colleagues studied mice previously infected with urinary tract bacteria, and treated the animals with low doses of the antibiotic ciprofloxacin. In mice that had cleared their infections before receiving the drug, 80% became reinfected. Another group of mice with a low level of infection had more bacteria in their urine after taking the antibiotics.

The antibiotic caused the bacteria to produce proteins that let them stick to bladder and kidney cells, making it easier for the pathogens to colonize these tissues.
mBio **6**, e00356-15 (2015)

MOLECULAR PATHOLOGY

Cancer spreads among clams

Outbreaks of leukaemia-like cancer in soft-shell clams may have originated in a single clam.

Mysterious cancers have been affecting clams and other marine bivalves in the United States and Europe since at least the 1970s. Stephen Goff at Columbia University in New York and his colleagues studied the DNA of cancerous and non-cancerous cells from several populations of

soft-shell clams (*Mya arenaria*) along the coast of the eastern United States. The DNA from cancerous cells did not match that of the hosts' other tissues, but the cancer cells were genetically similar to each other, suggesting that they arose from a single ancestor.

Only two other transmissible cancers are known, affecting dogs and Tasmanian devils. However, invertebrates may be particularly vulnerable because they lack a part of the vertebrate immune system that identifies foreign invading cells, the authors say.
Cell **161**, 255–263 (2015)

PHYSICS

Hot fluids act strangely in space

Boiling fluids behave differently in space and on Earth, suggesting that new approaches are needed to cool spacecraft in orbit.

Heat pipes suck excess heat away from laptop computers and other devices, and consist of a tube filled with liquid that evaporates at one end when heated. The vapour flows to the cool end, then condenses and returns to the other end. Joel Plawsky of Rensselaer Polytechnic Institute in Troy, New York, and his colleagues sent a heat-pipe experiment to the International Space Station (**pictured**), where the transparent, pentane-containing pipe was heated.

Surprisingly, the liquid did not rush away from the hot end as it does on Earth, but instead flooded the heated area. In zero gravity, capillary forces pulled liquid towards the hot end, whereas on Earth, gravity counteracts these forces.
Phys. Rev. Lett. **114**, 146105 (2015)

↻ **NATURE.COM**
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS
The news in brief

Ebola drug restart

Phase I trials of an experimental Ebola drug will restart after US regulators modified restrictions they had placed on the study. Tekmira Pharmaceuticals of Burnaby, Canada, said on 10 April that the US Food and Drug Administration (FDA) will allow the company to administer TKM-Ebola to a number of healthy people for up to one week. The FDA had halted the study in July 2014, requesting more information about how the drug works (see *Nature* **511**, 520; 2014). Although regulators later allowed use of the drug in patients infected with Ebola, testing higher doses in healthy people remains on hold.

Precision medicine

On 14 April, California launched a statewide US$3-million precision-medicine initiative to study how genomic, socio-economic, environmental, mobile and other forms of patient data can be combined to inform the development of drugs and the better practice of medicine. Hosted at the University of California, San Francisco, the initiative will be led by Atul Butte, director of its Institute for Computational Health Sciences. The effort follows the US Precision Medicine Initiative announced in January, a national project to collect data from one million people. See go.nature.com/2zelzo for more.

Stop-and-go scope

Organizers of the Thirty Meter Telescope on Mauna Kea in Hawaii will halt construction until at least 20 April, Hawaii's governor David Ige announced on 11 April. Last week, dozens of protesters were arrested for

DON PREISLER/UC DAVIS SCHOOL VET. MED.

Breeding programme to boost rare voles

An endangered population of California voles may soon be helped towards recovery by animals raised in captivity. Researchers at the University of California, Davis, announced on 10 April that a breeding programme for the Amargosa vole (*Microtus californicus scirpensis*; pictured) is preparing to release its first animals into the wild. The subspecies has been driven almost to extinction by loss of habitat and by climate change; only a few hundred are estimated to remain in the Mojave Desert marshes. The programme, started in July 2014 in collaboration with state and federal wildlife officials and the University of California, Berkeley, has grown from 20 to 90 captive voles. The researchers plan to release about two dozen animals into two desert marshes near Tecopa, and will track the voles using radio transmitters for up to a year.

trying to block building work on the mountain's summit. Many Native Hawaiians consider Mauna Kea to be sacred, and some have filed lawsuits against the project.

Disease control

The African Union and the United States signed an agreement on 13 April to create the African Centres for Disease Control and Prevention (CDC). The African CDC will launch later this year, beginning with a surveillance and response unit to assist in public-health emergencies on the continent. As part of the agreement, the US CDC will second two public-health experts to the African Union to act as long-term technical advisers, and will provide fellowships for ten African epidemiologists.

Research fraud

The US Office of Research Integrity has uncovered a series of data fabrications by neuroscientist Ryousuke Fujita. Fujita, formerly a postdoctoral researcher at Columbia University in New York City, had previously admitted to faking results in a retracted 2011 *Cell* paper that described the conversion of skin cells from people with Alzheimer's into neurons. The office's findings, released on 7 April, also reveal sample-size inflation and image manipulation in a 2013 *Nature* paper and in an unpublished manuscript. Fujita has agreed to exclude himself from federal research funding and from peer-review committees for agencies such as the US National Institutes of Health for three years.

Retraction request

Neuroscience researcher Teresita L. Briones will request the retraction of five publications as part of an agreement with the US Office of Research Integrity announced on 7 April. The office found that the former professor at Wayne State University in Detroit, Michigan, "intentionally, knowingly,

and recklessly" falsified and fabricated data related to studies of neuroinflammation, cognitive impairment and the accumulation of amyloid proteins in a rat model of brain injury. The faked results also affect three grant applications submitted to the US National Institutes of Health.

Psychiatry chief

The University of Minnesota in Minneapolis announced on 9 April the resignation of Charles Schulz, head of its psychiatry department. Schulz said that he wanted to focus on his medical practice and make way for new leadership. The university is currently reviewing and revamping its ethics policies for clinical research, after an external report found inadequate protections for human participants in psychiatric studies. Enrolment in all of the department's interventional drug trials have been suspended since March (see *Nature* http://doi.org/3nk; 2015).

FUNDING

Exascale computer

The US Department of Energy will spend US$200 million on a next-generation supercomputer for Argonne National Laboratory in Illinois, it announced on 9 April. The machine, to be called Aurora, uses an Intel high-performance computing system and is due to open for scientific research in 2018. The grant is the third and final in the energy department's push towards exascale computing, a milepost expected to be reached in the early 2020s (see *Nature* **515,** 324; 2014).

Transgenic tree

Brazilian regulators approved on 10 April the commercial use of a genetically modified eucalyptus species developed by biotechnology firm FuturaGene of Rehovot, Israel. The eucalyptus is engineered to grow faster and produce about 20% more wood (see *Nature* **512,** 357; 2014) than do conventional trees. Use of the plant could free up some industrial forest land, the company said; roughly 3.5 million hectares are currently occupied by eucalyptus plantations across Brazil (**pictured**). The decision paves the way for the world's first large-scale commercial deployment of a genetically modified tree.

Contract cool-off

Energy provider Southern Company in Atlanta, Georgia, confirmed last week that it will not be renewing its funding agreement with the Harvard–Smithsonian Center for Astrophysics in Cambridge, Massachusetts, when the agreement expires later this year. The centre and one of its researchers, climate-change sceptic Willie Soon, came under fire in February after documents revealed the terms of their earlier contracts with the company. In one case, they agreed to notify the company if disclosing it as a source of funding. See go.nature.com/khqcem for more.

TREND WATCH

Loss of tree cover has surged since 2011 in the boreal forests of Russia, Canada and Alaska, according to an analysis of satellite data released this month by the World Resources Institute in Washington DC (see go.nature.com/6izl4p). The authors suggest that recent spikes in forest fires, which vary greatly from year to year, could be to blame. In the long term, it is predicted that climate change could lead to more frequent and intense boreal wildfires in the twenty-first century.

BOREAL BREAKDOWN

Tree cover losses in northern boreal forests have spiked in recent years, but worldwide, losses in tropical regions continue to dominate.

Annual tree cover loss (millions of hectares)

- Tropical
- Boreal
- Temperate
- Subtropical

Trend lines represent moving averages of each year and the previous two years.

COMING UP

18–22 APRIL
Highlights at the annual meeting of the American Association for Cancer Research in Philadelphia, Pennsylvania, include developments in antibody–drug complexes and stem-cell cultures for drug testing.
go.nature.com/obebc4

20–23 APRIL
The Space Telescope Science Institute in Baltimore, Maryland, hosts Hubble's 25th Anniversary Symposium, where astronomers will share results from the telescope.
go.nature.com/tcuzoe

21–23 APRIL
Tsunami resilience and the future of earthquake early-warning systems are on the agenda at the annual meeting of the Seismological Society of America in Pasadena, California.
go.nature.com/17ifbo

BUSINESS

Eyes on natural gas

Oil-and-gas giant Royal Dutch Shell will take over the UK gas firm BG Group in a US$70-billion deal announced on 8 April. BG's natural-gas holdings are expected to give Shell a leg up in the fast-growing market for liquefied natural gas, a cleaner-burning alternative to coal for generating electricity and heating homes. Industry experts at Wood Mackenzie, an energy analysis firm headquartered in Edinburgh, UK, say that the combined company is on track to become the biggest seller of liquefied natural gas by 2018.
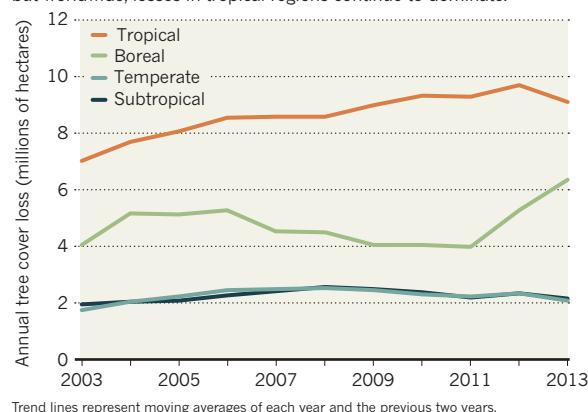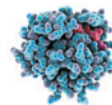
↻ **NATURE.COM**
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

USAF/GETTY

Experiments that use only a small number of animals are common, but might not give meaningful results.

MEDICAL RESEARCH

UK funders demand strong statistics for animal studies

Move addresses concerns that some experiments are not using enough animals.

BY DANIEL CRESSEY

Replace, refine, reduce: the 3 Rs of ethical animal research are widely accepted around the world. But now the message from UK funding agencies is that some experiments use too few animals, a problem that leads to wastage and low-quality results.

On 15 April, the research councils responsible for channelling government funding to scientists, and their umbrella group Research Councils UK, announced changes to their guidelines for animal experiments. Funding applicants must now show that their work will provide statistically robust results — not just explain how it is justified and set out the ethical implications — or risk having their grant application rejected.

The move aims to improve the quality of medical research, and will help to address widespread concerns that animals — mostly mice and rats — are being squandered in tiny studies that lack statistical power.

"If the study is underpowered your results are not going to be reliable," says Nathalie Percie du Sert, who works on experimental design at the National Centre for the Replacement, Refinement and Reduction (NC3Rs) of Animals in Research in London. "These animals are going to be wasted."

Researchers say that sample size is sometimes decided through historical precedent rather than solid statistics. There is also a lack of clarity: last year, an analysis of selected papers published in Nature or Public Library of Science journals describing animal experiments revealed that few reported the use of statistical tests to ▶

ANIMAL USE
In 2013, fundamental biology accounted for most animal experiments in the United Kingdom.

2 MILLION PROCEDURES*

Fundamental biological research
58%

Human medicine or dentistry
26%

Veterinary medicine
8%

Other
3%

Toxicological and related evaluations
5%

*Breeding procedures such as genetic modification not included.

▶ determine sample size, even though both publishing groups had endorsed guidelines to improve reporting standards (D. Baker *et al. PLoS Biol.* **12,** e1001756; 2014).

Animals feature in a wide range of experiments (see 'Animal use'), many of which are designed to test drugs before trials are done in people. The effects that researchers are looking for in these preclinical studies are often subtle, and 'power calculations' are needed to reveal the number of animals needed to show an effect. But an international academic partnership called the CAMARADES project (Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies), has shown that many animal studies are underpowered: studies in stroke, for example, are typically powered at between 30% and 50%, meaning that there is just a 30–50% chance of detecting a biological effect if it exists.

Malcolm Macleod, a neuroscientist at the University of Edinburgh, UK, blames, among other things, a lack of training and support in experimental design, as well as limited funds: animals are expensive to work with.

Some say that the pressure to 'reduce' may be one of the reasons for small experiments, but others counter that this is a misinterpretation of the 3 Rs because small experiments are ethically problematic if they have low statistical power.

The problem is not limited to Britain: last year, Francis Collins, director of the US National Institutes of Health (NIH), and Lawrence Tabak, NIH deputy director, warned about a lack of reproducibility in preclinical research and mentioned a dearth of sample-size calculations as one of the problems (see *Nature* **505,** 612–613; 2014).

The situation infuriates animal-welfare proponents. "It's completely unethical to use animals in studies that aren't properly designed," says Penny Hawkins, head of the research-animals department at the Royal Society for the Prevention of Cruelty to Animals in Southwater, UK.

Boosting the number of animals in specific experiments need not mean more animals are used overall because multiple small experiments can often be replaced by fewer, larger ones "One potential implication is we need to ask for money to do larger studies," says Marcus Munafò, a psychologist at the University of Bristol, UK.

Another way to increase sample sizes would be to link up researchers working on similar topics. Munafò notes that this is what geneticists now do for studies that require scanning a large number of genomes. "That template already exists," he says. "The question is, how do you initiate that cultural change?"

More immediately, du Sert is developing an online tool for the NC3Rs that will help researchers to design robust studies. "We're not blaming anyone for the way they were doing things before," she adds. "That was the practice at the time." ■ **SEE EDITORIAL P.263**

Canadians baulk at reforms to health–research agency

Biomedical–funding revamp threatens to marginalize under–represented researchers.

BY SARA REARDON

The biggest overhaul in the 15-year history of the Canadian Institutes of Health Research (CIHR) was meant to rescue biomedical researchers from the endless grant applications and Byzantine peer-review processes that had become a feature of the cash-strapped agency. "The research community was complaining bitterly," says Alain Beaudet, president of the CIHR in Ottawa. "They begged me to make changes."

But now that reality is kicking in, many researchers worry that the changes — which modify how grants are awarded, restructure advisory boards and reallocate the money funnelled through the 13 virtual institutes that comprise the CIHR — will marginalize some fields and hurt early-career researchers.

Beaudet says that the plans have been in place for some time, but many researchers — particularly those on the institutes' scientific advisory boards — complain that the CIHR has failed to communicate the changes adequately, and that the number of simultaneous reforms is overwhelming.

"We're a little bit stunned," says Gillian Einstein, a cognitive neuroscientist at the University of Toronto and chair of the board that advises the CIHR's Institute of Gender Health. "I'm not sure the groundwork was laid so we'd understand what was happening."

Each institute has its own advisory board with up to 12 members, and receives a dedicated allotment of about Can\$8.5 million (US\$6.7 million) from the CIHR's Can\$1-billion annual research budget. In the 2016 budget, these outlays will be cut in half, with the savings going into a common fund. To access this new funding source, institutes will have to work together to design cross-disciplinary initiatives that have extra support from a funding partner such as a charity, institution or company. Beaudet says that the CIHR will be responsible for finding many of these partners.

The CIHR also plans to eliminate most of the scientific advisory boards, leaving only three or four panels, which will advise several institutes each. An internal panel is still evaluating the plan, which would not take effect before April 2016. Nearly all of the advisory boards are protesting the changes. "If you're doing well and have some vision, and someone took half your toolset away, I'd say the rug was pulled out," says Anthony Jevnikar, a nephrologist at Western University in London, Ontario, who chairs the advisory board for the Institute of Infection and Immunity.

BAR TO ENTRY
Feathers are also being ruffled by changes to the CIHR's system for awarding grants to proposals submitted by researchers. In July,

the agency plans to hand out the first set of awards under a pilot system that divides about half of its research budget between two mechanisms. One of these, the Foundation Scheme, gives seven years of guaranteed funding to established researchers and five years to early-career investigators. Grant recipients can use the money for any project, but are barred from receiving other CIHR funding. The second mechanism, the Project Scheme, awards smaller grants for specified work over a shorter period.

But researchers who have been reviewing the first set of applications under the new system see potential problems, particularly for early-career researchers, who often have difficulty showing enough preliminary data to justify specific projects or enough of a track record to win an open-ended grant. New investigators submitted about 40% of the 1,366 grant applications for the Foundation Scheme's pilot round, but they were involved with less than 20% of the 467 applications that made it through the first phase of peer review. "Young researchers are left out in the cold," says Jim Woodgett, a molecular biologist at Mount Sinai Hospital in Toronto.

Some institutes also feel imperilled by the changes. Researchers supported by the Institute of Aboriginal Peoples' Health (IAPH) say that they have few funding options outside the CIHR, and would not find it easy to interest external partners in providing support so that they could receive money through the cross-disciplinary common fund. Their field is relatively new and they are under-represented among public-health researchers, so they feel disadvantaged if they have to compete against other institutes for money and for spots on an advisory board that will also oversee other institutes. "We're losing our distinctive voice," says Frederic Wien, a sociologist at Dalhousie University in Halifax who studies aboriginal health.

Such concerns are exactly why the reforms are taking place, says Beaudet: "There were not enough collaborations between institutes." For instance, he says, the other 12 institutes assumed that they did not need to worry about aboriginal peoples' health, because the IAPH would cover all relevant research. The other institutes' inattention to indigenous peoples' health is a huge problem, Beaudet adds.

Wien says that the CIHR has not been responsive to complaints over the past several years. He and others are also concerned that the agency might eliminate some institutes altogether. The 13 divisions have existed since the CIHR was founded, but Beaudet says that, by law, external and internal panels must review the institutes every five years; it has always been possible that some could be eliminated. ∎

The increasing sophistication of 3D printing is shown in an ear that melds biological and electronic parts.

MATERIALS

Printed body parts come alive

Conference on 3D printing features made-to-order bones, and organs built using cells as 'ink'.

BY HEIDI LEDFORD

The advent of three-dimensional (3D) printing has generated a swell of interest in artificial organs meant to replace, or even enhance, human machinery.

Printed organs, such as a prototype outer ear developed by researchers at Princeton University in New Jersey and Johns Hopkins University in Baltimore, Maryland, will be on the agenda at the Inside 3D Printing conference in New York on 15–17 April. The ear is printed from a range of materials: a hydrogel to form an ear-shaped scaffold, cells that will grow to form cartilage, and silver nanoparticles to form an antenna (M. S. Mannoor *et al. Nano Lett.* **13,** 2634–2639; 2013). The device is just one example of the increasing versatility of 3D printing.

The New York meeting, which bills itself as the largest event in the industry, will have plenty of widgets and novelties on display. But it will also feature serious discussions on the emerging market for printed body parts.

That business is

NATURE.COM
For more printed parts, see:
go.nature.com/qsy61w

currently focused on titanium replacement hip joints, which can be tailored to fit individual people, and made-to-order polymer bones to reconstruct damaged skulls and fingers. Printed body parts brought in US$537 million last year, up about 30% on the previous year, says Terry Wohlers, president of Wohlers Associates, a business consultancy firm in Fort Collins, Colorado, that specializes in 3D printing.

Scientists are looking ahead to radical emerging technologies that use live cells as 'ink', assembling them layer-by-layer into rudimentary tissues, says Jennifer Lewis, a bioengineer at Harvard University in Cambridge, Massachusetts. Bioprinting firm Organovo of San Diego, California, already sells such tissues to researchers aiming to test experimental drugs for toxicity to liver cells. The company's next step will be to provide printed tissue patches to repair damaged livers in humans, says Organovo's chief executive, Keith Murphy.

Lewis hesitates to say that 3D printing will ever yield whole organs to relieve the shortage of kidneys and livers available for transplant. "I would love for that to be true," she says. "But these are highly complicated architectures." ∎

Iran's Fordow nuclear-enrichment plant could be converted into an international physics laboratory.

IRAN NEGOTIATIONS

Hope for science in fallout of nuclear deal

Iranian physicists excited at prospects of a new physics lab and greater collaboration with the rest of the world.

BY DECLAN BUTLER

The preliminary deal agreed between six world powers and Iran over its nuclear programme has been hailed as an opportunity to end years of global tension, prevent a nuclear arms race in the Middle East and ease sanctions that have crippled Iran's economy. Iranian science may also be a winner.

Negotiators must still resolve many outstanding issues before the 30 June deadline for a formal, written deal (see 'Challenges to a formal deal'). But researchers in and outside Iran are cautiously optimistic about an intriguing spin-off of the agreement — a proposal to convert the Fordow uranium-enrichment plant into an international physics laboratory, as well as opportunities for collaborations that will arise if sanctions are eased.

"The option that Fordow may be transferred into a physics-research facility is certainly exciting to the physics community of Iran," says Shahin Rouhani, a physicist at the Institute for Research in Fundamental Sciences (IPM) in Tehran and president of the Physics Society of Iran. "What will actually happen is, of course, dependent on the final agreement in a few months' time, and the exact infrastructure available in Fordow," he adds.

Buried beneath a mountain, the Fordow facility concerns the United States and its allies because it would be difficult to destroy, so the proposal to turn it into a lab seems above all a diplomatic device. "I think that this is primarily a way to bridge over the Ayatollah's requirement that no Iranian nuclear facility be shut down and the US requirement that enrichment of uranium stop there," says Frank von Hippel, a nuclear-weapons and non-proliferation physicist at Princeton University in New Jersey.

According to the proposed deal, some of the uranium-enrichment centrifuges at the Fordow site would be repurposed to produce isotopes such as molybdenum-99, which is widely required for medical imaging (see go.nature.com/jafnpt). Rüdiger Voss, head of international relations at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, says that such a capability could help to stem a global shortage of these isotopes.

Other parts of the underground site would house experimental physics facilities; the usefulness of this would depend heavily on the nature of the facilities, which are vague right now, says Voss. Physicist Ernest Moniz, the US energy secretary who is the nation's lead scientific negotiator on the agreement, has mentioned the possibility of installing a particle accelerator there. Iranian physicists are already on the case too, says Reza Mansouri, an astronomer at the IPM and a former deputy science minister. The Physics Society of Iran intends to write to politicians to ask to be involved in the choice of any future projects, he says, and Rouhani plans for the society to set up a working group to examine the possibilities for exploiting Fordow. He cites the construction of a neutrino detector as a possibility; Mansouri suggests research on cosmic rays.

Iran has a vibrant physics community and is already home to facilities such as the Iranian Light Source Facility in Qazvin, northwest of Tehran, which provides intense beams of X-rays for research in many fields. Iranian physicists also have many international collaborators, for example through the country's participation in experiments at CERN, and so are well poised to discuss any plans with colleagues abroad. "I can only be open to the initiative," says Patrick Fassnacht, who is in charge of international relations with Iran at CERN.

The negotiations also include easing nuclear-related sanctions, which have affected Iran's ability to do research and to collaborate with foreign scientists, says Hamid Javadi. He is a member of the council of the Iranian-American Physicists group, a body set up in 2007 to represent Iranian members of the American Physical Society. Foreign scientists often avoid contact with their Iranian peers for fear of falling foul of the tough sanction laws, he says. Iranian scientists wishing to travel abroad have also had difficulty obtaining visas.

Furthermore, the sanctions have made experimental equipment and journal subscriptions expensive for Iranian researchers, says Warren Pickett, a physicist at the University of California, Davis, who has promoted science diplomacy with Iran through visits (W. E. Pickett *et al. Nature Phys.* **10,** 465–467; 2014), and whose university last year agreed to collaborate

with Sharif University of Technology in Tehran. More perniciously, international tensions have often driven a wedge between foreign and Iranian researchers, he says: "When I described my visit to Iran, some colleagues would seem to roll their eyes in a 'why would you go there?' fashion."

He adds: "Introducing a large country of 75 million people back into the international community would be a great breakthrough."

Eased sanctions would free up Iran's existing collaborations, too. "Life is not easy for our Iranian friends," says Fassnacht. CERN itself has had to be careful not to inadvertently contravene sanctions when dealing with Iran, he says, such as working with people blacklisted for their links to the country's nuclear programme. Swiss banks have also been reluctant to accept Iran's payment of dues to CERN, although CERN finally found a bank willing to do so, he adds.

The SESAME synchrotron being built near Amman, Jordan, with a goal of promoting peace between Middle Eastern nations, as well as particle physics, has faced similar bank problems, says Christopher Llewellyn-Smith, director of energy research at the University of Oxford, UK, and president of the SESAME council. "It will be a real shot in the arm for SESAME, as it will allow Iran to start paying again and pay debts which have accumulated since sanctions began," he says.

Even more broadly, the negotiations signal a readiness for dialogue. "It's immensely important," says Mansouri, "that Iran, the US and other countries have learnt to talk with each other with rationality." ■ **SEE EDITORIAL P.263**

Additional reporting by Davide Castelvecchi

NUCLEAR SCIENCE
Challenges to a formal deal

US and Soviet leaders relied on physicists to work out nuclear-weapons reductions and verification procedures during the cold war. Similarly, negotiators now working on a formal deal on Iran's nuclear programme are looking to scientists to provide confidence in its technical underpinnings. Here are three nuclear capabilities that a final agreement will need to address.

BREAKOUT Central to the deal is the concern that Iran could quickly divert its nuclear programme — which it claims is for peaceful purposes — to produce the highly enriched uranium or weapons-grade plutonium needed to build a bomb, an event known as 'breakout'. The preliminary deal requires that Iran reduce the number of operating centrifuges from 19,000 to 5,060 and its stockpile of low-enriched uranium from 10,000 kilograms to 300 kg. Under this scenario, it would take at least a year after a breakout to produce the uranium needed for a bomb, enough time for intervention. The framework agreement also stipulates that the core of a heavy-water nuclear reactor at Arak be replaced with one that generates less plutonium in its spent fuel — and that all spent fuel be sent out of the country.

SNEAK-OUT Under the framework agreement, for the next 25 years the International Atomic Energy Agency (IAEA), headquartered in Vienna, would be given unprecedented powers to inspect any part of Iran's nuclear-fuel cycle. It would also have the right to investigate the possibility of 'sneak-out' — undeclared sites carrying out uranium enrichment or other activities that could result in nuclear weapons. The agency's inspections would use satellite imagery, searches for equipment and environmental sampling to check whether highly enriched uranium has been used at a site. But the agreement is currently much less detailed when it comes to sneak-out than for breakout, and Iran has in the past hidden enrichment plants from the IAEA.

WEAPONS RESEARCH Perhaps the thorniest issue is military nuclear research. If inspectors had access to the nation's Parchin military site, where work on the development of nuclear weapons is alleged to have taken place, they could look for evidence of the testing of nuclear-weapons components. But the deal as laid out does not touch on what powers the IAEA would have to inspect military sites, and, unsurprisingly, Iran has in the past refused the IAEA access to Parchin. Satellite images suggest that Iran has tried to conceal previous nuclear-weapons research at the site from any future IAEA inspection. **D.B.**

GENDER

Leading scientists favour women in tenure–track hiring test

US science and engineering professors preferred female job candidates by two to one.

BY BOER DENG

Universities in the United States employ many more male scientists than female ones. Men are paid more, and in fields such as mathematics, engineering and economics, they hold the majority of top-level jobs.

But in a sign of progress, a 13 April study finds that faculty members prefer female candidates for tenure-track jobs in science and engineering — by a ratio of two to one. That result, based on experiments involving hypothetical job seekers, held true regardless of the hirer's gender, department, career status or university type, researchers report in the *Proceedings of the National Academy of Sciences*[1].

"We were shocked," says Wendy Williams, a psychologist at Cornell University in Ithaca, New York, and a co-author of the study. With fellow Cornell psychologist Stephen Ceci, she surveyed 873 tenure-track faculty members in biology, psychology, economics and engineering at 371 US universities. One experiment presented participants with three hypothetical job candidates, of which two were identical except for their gender. Another experiment added descriptions of marital and parental status, to test whether underlying assumptions about gender choices affected hiring. "You don't frequently see that level of attention and sophistication" in statistical analysis, says Robert Santos, vice-president of the American Statistical Association in Alexandria, Virginia.

Nothing seemed to sway study participants' preference for female job candidates. The authors say that this is interesting given their previous finding that a relatively low percentage of female PhDs in the social and biological sciences secure academic positions — in part ▶

because they are less likely than men to apply for these jobs. Other research suggests that in the physical sciences, women and men are just as likely to secure a tenure-track position within five years of earning a PhD.

There are more signs that science is inching towards gender equality. In February, a study[2] in the journal *Frontiers in Psychology* reported that US women and men with bachelor's degrees in science, engineering and mathematics go on to receive doctoral degrees at roughly the same rate.

Nancy Hopkins, a biologist at the Massachusetts Institute of Technology in Cambridge, argues that the news is not as good as it seems. Women in academic science still face gender-related obstacles before they reach the point of applying for tenure-track jobs, she says.

In the biological sciences, for example, most elite US labs are headed by men. These principal investigators hire more male postdoctoral researchers than female ones[3] — despite the fact that women receive the majority of biology doctorates. Postdocs from such elite labs also tend to be chosen for assistant-professor positions, perpetuating the cycle[3]. Other studies have found that individual faculty members of both genders view female students as less competent than their male counterparts when judging qualifications for junior positions in a lab[4].

Virginia Valian, a psychologist at Hunter College in New York who studies gender equity, says the study's main findings are not surprising. But, she says, "there is a valid concern that progress will be over-interpreted."

Asked about the doubt that has greeted the study, Williams argues that "people find it hard to accept when there's change, even for the better." But she does not dispute that bias may still undermine the prospects of women in science. She and Ceci are now examining women's chances of advancement at other points in their scientific careers, on the basis of data from other nationally representative surveys. ∎

1. Williams, W. M. & Ceci, S. J. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1418878112 (2015).
2. Miller, D. I. & Wai, J. *Front. Psychol.* **6,** 37 (2015).
3. Sheltzer, J. M. & Smith, J. C. *Proc. Natl Acad. Sci. USA* **111,** 10107–10112 (2014).
4. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelman, J. *Proc. Natl Acad. Sci. USA* **109,** 16474–16479 (2012).

Computer systems have been prone to error since the early days.

USAF/GETTY

REPRODUCIBILITY

Journal buoys code–review push

Nature Biotechnology *asks peer reviewers to check accessibility of software used in computational studies.*

BY ERIKA CHECK HAYDEN

The finding seemed counterintuitive: warming in North America was driving plant species to lower elevations — not towards higher, cooler climes, as ecologists had long predicted. But the research published in *Global Change Biology* indeed turned out to be wrong. In February, the journal retracted the paper after its intriguing conclusion was found to be the result of errant software code[1].

Worried about a rising tide of results that fail to measure up, journals are starting to take action. In the latest such move, *Nature Biotechnology* announced on 7 April a plan to prevent such embarrassing episodes in its pages (*Nature Biotechnol.* **33,** 319; 2015). Its peer reviewers will now be asked to assess the availability of documentation and algorithms used in computational analyses, not just the description of the work. The journal is also exploring whether peer reviewers can test complex code using services such as Docker, a piece of software that allows study authors to create a shareable representation of their computing environment.

Researchers say that such measures are badly needed. They note that the increasing size of

---

MORE ONLINE

ASTEROID MISSION
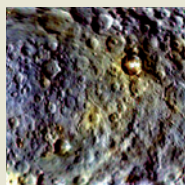
Latest images from Ceres show craters and puzzling 'bright spots' go.nature.com/j42dwj

MORE STORIES

● γ-ray method flags up nuclear stash in shipping cargo go.nature.com/p9mwm3
● *Nature* readers have their say on science's postdoc crisis go.nature.com/7rffu8
● How to tell male and female dinosaurs apart go.nature.com/vioq8u

NATURE PODCAST

Self-experimentation in science and the wedding rings that went to Hubble nature.com/nature/podcast

NASA/JPL-CALTECH/UCLA/MPS/DLR/IDA

data sets and complexity of analysis software makes errors harder to detect. "This is a big step forward," says Ciera Martinez, a plant biologist at the University of California, Davis. "A large journal focusing on reproducibility is desperately needed." Computational experts often raise issues about code quality or availability during peer review, she adds, but such concerns are often ignored because many journals do not require examination of code.

The result can be errors or irregularities that lead to retractions, corrections and divisive debates. In announcing its policy, *Nature Biotechnology* cited two of its studies that were called into question by scientists who could not replicate the conclusions. Both papers[2,3] had reported new methods for analysing connections within networks, but neither provided sufficient documentation of their tools or approach. The journal has now published more information about how software was used in each analysis.

"We are simply seeking to make our editorial evaluation of computational tools more consistent," says *Nature Biotechnology* editor Andrew Marshall, who adds that other journals that publish computational-biology research have taken similar steps.

But several issues complicate the drive for software reproducibility. One is the difficulty of finding qualified reviewers for papers in disciplines that cross departmental boundaries. "The research is collaborative, but the review process is stuck in a disciplinary mindset," says Lior Pachter, a computational biologist at the University of California, Berkeley.

Another is social: there is no etiquette governing how those who wish to replicate results should behave towards those whose work they examine. If authors of erroneous studies face public embarrassment and shaming, that can discourage other researchers from submitting to the same scrutiny. "It's like taking your clothes off; you don't want to be embarrassed by someone pointing at you because you have a lot of body hair," says Ben Marwick, an archaeologist at the University of Washington in Seattle.

Mindful of such concerns, advocates of software reproducibility are placing less emphasis on publications. Instead they argue that published tools should be able to be used by other researchers. They say that this approach acknowledges the iterative nature of science.

"When we say 'open science' or 'open research', it's not just about accessibility and availability of content or material," says Kaitlin Thaney, director of the non-profit Mozilla Science Lab in New York. "It's taking it one step further to think about use and reuse, so someone can carry that forward."

An increasing number of initiatives aim to encourage scientists to ensure that their software is replicable. Courses run by organizations such as the non-profit Software Carpentry Foundation teach the value of writing and sharing solid scientific code, as well as the principles of constructing it. Software packages such as iPython and knitr make it easier to document code creation transparently and in its research context. The Mozilla Science Lab has experimented with training researchers in the scientific-coding process, and universities such as the University of California, Berkeley, are creating courses that train graduate students to code in a way that advances the cause of open and reproducible science.

The cause has been slow to catch on in the upper echelons of research. But those pushing for great replicability hope that a combination of incentives could begin to make a difference. Measures aimed at the publication process, such as those announced by *Nature Biotechnology*, will hit home for many researchers. Others may be lured by the notion that replicable work is more likely to stand the test of time. "The incentive for me, as a young researcher, is simple," says Martinez. "Better science." ∎

1. Harsch, M. A. & Hille Ris Lambers, J. *Glob. Change Biol.* **21,** 1376 (2015).
2. Barzel, B. & Barabási, A.-L. *Nature Biotechnol.* **31,** 720–725 (2013).
3. Feizi, S., Marbach, D., Médard, M. & Kellis, M. *Nature Biotechnol.* **31,** 726–733 (2013).

THE RAS RENAISSANCE

Thirty years of pursuit have failed to yield a drug to take on one of the deadliest families of cancer-causing proteins. Now some researchers are taking another shot.

BY HEIDI LEDFORD

When Stephen Fesik left the pharmaceutical industry to launch an academic drug-discovery laboratory, he drew up a wanted list of five of the most important cancer-causing proteins known to science. These proteins drive tumour growth but have proved to be a nightmare for drug developers: they are too smooth, too floppy or otherwise too finicky for drugs to bind to and block. In the parlance of the field, they are 'undruggable'.

One of the first culprits that Fesik added to his list was a protein family called Ras. For more than 30 years, it has been known that mutations in the genes that encode Ras proteins are among the most powerful cancer drivers. Ras mutations are found in some of the most aggressive and deadly cancers, including up to 25% of lung tumours and about 90% of pancreatic tumours. And for some advanced

cancers, tumours with Ras mutations are associated with earlier deaths than tumours without them.

Decades of research have yet to yield a drug that can safely curb Ras activity. Past failures have driven researchers from the field and forced pharmaceutical companies to abandon advanced projects. But Fesik's laboratory at Vanderbilt University in Nashville, Tennessee, and a handful of other teams have set their sights anew on the proteins. They are armed with improved technology and a better understanding of how Ras proteins work. Last year, the US National Cancer Institute launched the Ras Initiative, a US$10-million-a-year effort to find new ways to tackle Ras-driven cancers. And researchers are already uncovering compounds that, with tweaking, could eventually yield the first drugs to target Ras proteins.

Researchers are mindful that they still have many hurdles to jump. "You have to have a lot of respect for Ras," says Troy Wilson, president of Wellspring Biosciences, a company in La Jolla, California, that launched in 2012 with its sights set on Ras. "It is not to be underestimated. But it's also one of the most important oncogenes in cancer."

Advocates of this Ras renaissance say that any signs of success could provide lessons on how to target other important proteins that are deemed to be undruggable. Just because people assume Ras proteins are too difficult to target does not mean that scientists should give up, says Channing Der, a cancer researcher at the University of North Carolina at Chapel Hill. "Dogma is a moving target."

HIGH-HANGING FRUIT

In 1982, Der's team was one of the first to show that mutations in human genes encoding Ras proteins can cause cancer[1]. This finding marked the culmination of a hunt for oncogenes — genes that can drive cancer — in the human genome. They had previously only been described in viruses and animal models.

The discovery laid the foundation for the modern cancer-research juggernaut, with its emphasis on tracking genetic mutations and mapping altered molecular pathways. It also prompted hopes of finding drugs that would target oncogenes and cure some cancers.

The following years were filled with discovery. It became clear that humans produce three highly similar Ras proteins and that these are activated when cells need to proliferate (to replace damaged tissue, for example). Signals from outside the cell switch Ras to an 'on' state, in which it is bound to a molecule called GTP. Cancer-causing forms of Ras proteins have a disabled 'off' switch and cannot properly process the GTP. So it seemed logical to search for drugs that could interfere with GTP binding to stop mutant Ras.

But as the understanding of Ras biochemistry grew, so too did a sense of pessimism. The family's affinity for GTP turned out to be extraordinarily high, and finding another compound that could block GTP's access seemed impossible. Ras proteins also work by interacting with other proteins, but small-molecule drugs that are able to get inside cells are often too small to cordon off the wide surface area usually involved in protein–protein interactions. (Antibodies can make excellent

> ## "PEOPLE SAID, 'NOBODY HAS DONE ANYTHING IN THE FIELD FOR TEN YEARS. LET'S DO SOMETHING'."

drugs and can mask a large area on their targets, but most do not penetrate cell membranes.)

Ras structures offered more reasons for concern. Drug developers look at a protein's shape to gauge the likelihood of finding a compound that will bind to a critical site. They like to see a protein with deep pockets that a drug can slip into and bind with multiple points of contact. However, Ras proteins are relatively smooth.

Twenty years ago, researchers thought they had the problem solved. To function, Ras proteins need to latch on to the inside of the cell membrane through a fatty tail. That tail is added by farnesyl transferase — an enzyme that is more amenable to drug targeting than Ras proteins. So the idea was to hobble Ras activity by finding drugs that inhibit farnesyl transferase.

At first, it looked like a winning strategy. Farnesyl transferase inhibitors damped down cell proliferation in mice and human cancer cells[2]. By the early 2000s, at least six pharmaceutical companies were racing to bring the drugs to market. Many abandoned other Ras-related projects because they thought the Ras problem was solved, says chemist Herbert Waldmann of the Max Planck Institute of Molecular Physiology in Dortmund, Germany. "The whole field took a deep breath and waited," he says.

The wait ended with one of the biggest disappointments in pharmaceutical history. One by one, the drugs failed in human clinical trials. Der, who was still studying Ras at the time, says that the episode taught him, and everyone else, an important lesson about Ras biology.

The three forms of human Ras are nearly identical in terms of structure and amino-acid sequence. Researchers assumed that their functions would be similar too. Most of the tools used to study Ras proteins — cell cultures, transgenic mice and antibodies — were developed using H-Ras, which was easier to work

with than the other forms. "All of us, including myself, thought why bother studying the other ones when we can just learn all about H-Ras," says Der. "Unfortunately, a lot of money was spent on that misconception."

It turned out that the other two forms of Ras in humans — K-Ras and N-Ras — are much more important in cancer, and the cell has a contingency plan in place to keep them working. In the absence of a farnesyl tail, another enzyme is able to tack on a different fatty tail, rendering the experimental drugs useless.

The Ras field was scarred by this episode, and it took some time before researchers were willing to give the proteins another look. But about a decade later, they started coming back. "All of a sudden people turned around and said, 'Hey, this is still one of the most important targets in oncology. Nobody has done anything in the field for ten years. Let's do something,'" says Waldmann. This time, researchers took a fresh approach by looking for weaknesses in Ras-driven tumours.

One such weakness is 'synthetic lethality'. When Ras proteins are in overdrive, cancer cells often become dependent on other molecular pathways for survival. Blocking these other pathways might not affect normal cells, but it kills Ras-driven tumour cells. Laboratories set about screening for the synthetic-lethal partners of mutated genes encoding Ras, with the idea that targeting them would kill cancer cells but leave normal cells unaffected.

The result was a wave of papers reporting possible new targets — followed closely by another wave of reports that the synthetic-lethal results were irreproducible[3]. Last October, William Sellers, Global Head of Oncology at the Swiss drug maker Novartis, reported at a conference that his team had tried and failed to reproduce the most prominent published Ras synthetic-lethal findings. Changes in context, such as the cell type used or specific screening conditions, could easily change the outcome of the experiment, says Julian Downward, a cancer researcher at the Francis Crick Institute in London. Researchers are still sifting through the results to find targets that hold up, but Downward is doubtful that the efforts will bear fruit. "Everyone seems to get something different from those experiments," he says. "I suspect these are not going to be the most robust targets."

TAILORED TO FIT

With the disappointment of the synthetic-lethal approach fresh in their minds, several researchers have been looking to target Ras itself (see 'Ras attack'). "We decided you have to go to Ras directly," says Brent Stockwell, a chemical biologist at Columbia University in New York.

Improvements made during the past five years in computer modelling and in ways of screening for drug compounds offer fresh hope for targeting the smooth, unpocketed terrain

RAS ATTACK

Ras proteins have proved devilishly hard to make drugs against. They have a relatively smooth surface with few pockets where a molecule might bind tightly.

Drugs that prevent the addition of an important fatty tail to Ras proteins failed in the clinic.
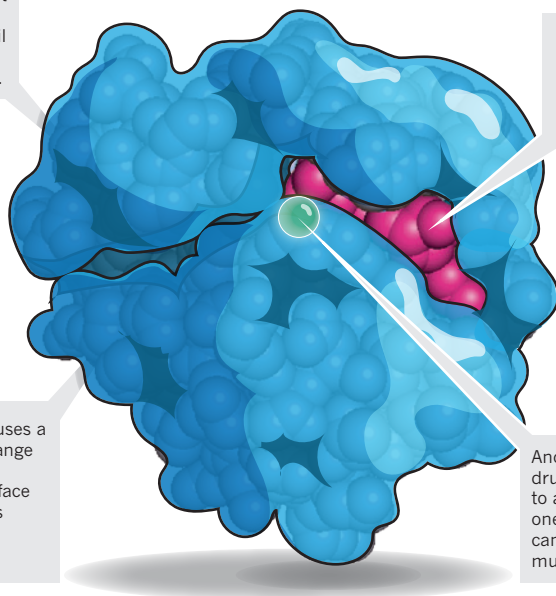
Ras proteins bind rapidly and tightly to GTP (in red), making it difficult to block the interaction.

One drug lead causes a Ras protein to change shape, forming a pocket on the surface of the protein. It is being refined to improve binding.

Another experimental drug binds irreversibly to a cysteine found on one of the most prevalent cancer-associated mutations in Ras.

of Ras proteins, Stockwell says. Researchers are now better able to predict the affinity of small molecules for proteins, for example, and have a better understanding of protein dynamics.

Stockwell's team is capitalizing on this to design small molecules that are tailored to the surface of Ras proteins — first in the computer, and then in the laboratory. "Maybe for these proteins, you're just not going to find the right solution anywhere out there in the world," Stockwell says. "You've just got to make it."

Fesik is also building new drugs, but starting from a library of existing compounds. In his former career at Abbott Laboratories in Abbott Park, Illinois, Fesik devised ways to disrupt interactions between proteins by piecing together fragments of compounds that bind, however weakly, to the target. The result is a large, novel compound that is unlikely to be found in the standard chemical libraries used to hunt for drugs.

Fesik likens the technique, called fragment-based screening, to constructing a key to fit a lock by cutting one notch at a time. "Eventually you combine all the notches," he says. "The compound has never been made before and yet you find it because you're building it up slowly and tailoring it to your protein."

Fesik's lab and his industry collaborators have found more than 130 molecules that bind weakly to K-Ras[4]. The compounds induce a change in the protein's structure, opening up a binding pocket in the process. The team is now trying to add on other fragments to improve the fit — in effect, the second notch in the key. Der notes that Fesik built a reputation for drugging the undruggable in industry before he left to pursue an academic career. "If anyone is going to do it, it is Fesik," he says.

Others are looking more closely at exploiting specific mutations within K-Ras. Although there are many different cancer-associated mutations in the gene that encodes it, just three are responsible for the vast majority of Ras-driven cancers. Each of these yields an enzyme with slightly different behaviour, says Der. "If we begin to think about different mutations as having different personalities, those different personalities may open up unique vulnerabilities," he says.

Kevan Shokat, a chemical biologist at the University of California, San Francisco, joined the Ras hunt six years ago. In 2013, he reported a compound that targets a K-Ras mutation known as G12C (ref. 5). The mutation, which is found in 20% of lung cancers, replaces the amino acid glycine with cysteine, which readily reacts with other molecules. Shokat's compound exploits the reactive cysteine and binds to it irreversibly. The inhibitor will require additional tinkering before it can be used in human patients but, as the first drug candidate that truly binds directly to Ras, it has generated a tremendous amount of excitement, says Downward. "It has re-energized the whole area," he says.

Shokat says he has long thought that a mutation-specific approach might work, but he hesitated to pursue it in his laboratory until recently. Drug developers were afraid of drugs that seize upon their target and do not come off, he says, because they seemed more likely to have unanticipated reactions with other proteins in the body. But several successful drugs, such as the lymphoma and myeloma drug ibrutinib, have recently been found to bind irreversibly to their targets.

Meanwhile, pharmaceutical companies are increasingly open to the idea of developing drugs that work in subsets of patients with cancer who carry specific mutations. "There won't be one drug that will work for every K-Ras patient," predicts Timothy Burns, a cancer researcher at the University of Pittsburgh in Pennsylvania.

Fesik says that the solutions to Ras's puzzles, whatever they are, will probably emerge from academic institutions. He left pharma in part because he loved the pursuit of important targets, regardless of how easy or hard they are to hit. Chasing an undruggable protein can be difficult to justify in industry, where scientific interest must often take a backseat to the near-term potential for profit. "Most pharma companies don't want to take the risk to go after these undruggable targets, and if they do, it's temporary," he says.

Bridges are forming, however. Fesik's laboratory has partnered with the German pharmaceutical company Boehringer Ingelheim to evaluate its first-generation Ras-binding drug. And Shokat co-founded Wellspring Biosciences to bring his inhibitor to market. The work soon won support from Janssen Biotech of Horsham, Pennsylvania.

The efforts are getting government attention as well. The multimillion-dollar Ras Initiative is supporting the development of tools and basic research on Ras protein structures to aid drug discovery, says Frank McCormick, a cancer researcher at the University of California in San Francisco and co-director of the project. "We are trying to de-risk Ras as a target so that others will jump back in the ring and have another shot," he says.

For years, the pharmaceutical industry has pursued low-hanging fruit in a different category of proteins called kinases, McCormick says. Those were easier to target, and yielded many useful cancer drugs. But that wave is starting to subside, he argues, and it is time to focus on the higher-hanging fruit: tougher targets, such as Ras proteins, that are known to be crucially important.

Stockwell says he hopes that the recent revival of research on Ras proteins could inspire scientists studying other intractable targets. "If there is some success there, maybe that excitement will extend to other targets," he says. "If we really want to impact disease, there's this vast space of additional targets that have never been mined." ■

1. Der, C. J., Krontiris, T. G. & Cooper, G. M. *Proc. Natl Acad. Sci. USA* **79,** 3637–3640 (1982).
2. Appels, N. M. G. M., Beijnen, J. H. & Schellens, J. H. M. *Oncologist* **10,** 565–578 (2005).
3. de la Cruz, F. F., Gapp, B. V. & Nijman, S. M. B. *Annu. Rev. Pharmacol. Toxicol.* **55,** 513–531 (2015).
4. Sun, Q. et al. *Angew. Chem. Int. Ed. Engl.* **51,** 6140–6143 (2012).
5. Ostrem, J. M, Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. *Nature* **503,** 548–551 (2013).
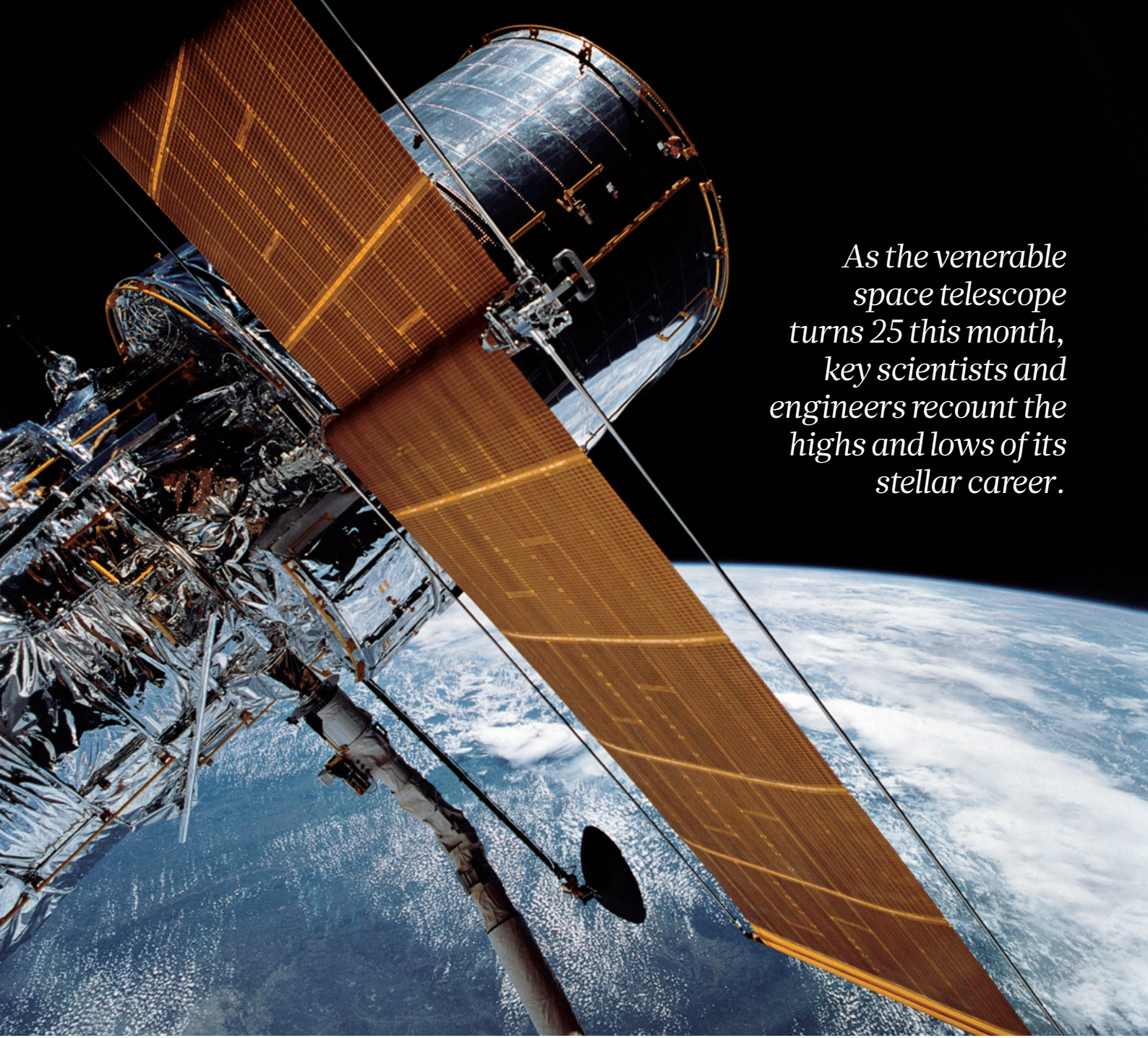
As the venerable space telescope turns 25 this month, key scientists and engineers recount the highs and lows of its stellar career.

VOICES OF HUBBLE

BY ALEXANDRA WITZE

NASA

When the Hubble Space Telescope blasted into space on 24 April 1990, it promised astronomers an unprecedented view of the Universe, free from the blurring effects of Earth's atmosphere.

But Hubble's quarter-century in orbit has never gone according to plan. The telescope — a joint venture between NASA and the European Space Agency (ESA) — faced a crippling flaw after launch that required astronauts to fly up and fix it. Later, problems with Hubble and NASA's shuttle programme left the telescope's future in jeopardy.

Through it all, Hubble emerged as the world's foremost astronomical observatory. Conceived by astronomer Lyman Spitzer in the 1940s, the telescope has led to fundamental discoveries, revealing for instance that the furthest reaches of the Universe are full of galaxies and that dark energy is pushing the cosmos apart at an ever faster rate. Its stunning images have transformed scientific understanding of the Universe and become wildly popular.

Here, *Nature* tells the story of Hubble through the words of some of its key players, beginning in 1972. At that time, the space telescope was little more than a set of engineering drawings.

"THE MONTHS IMMEDIATELY AFTER LAUNCH WERE JUST A NIGHTMARE."

Workers inspect Hubble's 2.4-metre main mirror in 1984.

ROBERT O'DELL, FORMER HUBBLE PROJECT SCIENTIST: I was told it would not take very long to build it. But I went in with my eyes wide open.

I could see that building Hubble was going to be the future. It was a chance to lead and influence the development of what I thought, even then, would be the most important telescope of my generation.

JEAN OLIVIER, FORMER HUBBLE CHIEF ENGINEER: Hubble was a proving ground for many technologies. Things you would think would be low-tech, like designing latches, evolved into a major problem. We kept uncovering more and more challenges.

It got to be such a long programme that I began to think it's not real life, it's a game — and one day they're going to say: "We're just kidding, we wanted to see how much you could take."

O'DELL: The lowest period was when it was becoming clear that we couldn't afford to do everything that we wanted to. This was right in the early hardware phase. I proposed that we would initially launch Hubble without all the instruments that were being developed. I proposed that out of desperation because people were actually saying we were going to cancel the programme unless you significantly reduce the costs. The lowest day for me was being chewed out in NASA headquarters for not standing up for the science of the project.

Hubble finally soared into orbit in 1990 aboard the space shuttle *Discovery*. But when the first image came back, it was blurry owing to a flaw known as spherical aberration.

SANDRA FABER, ASTRONOMER, UNIVERSITY OF CALIFORNIA, SANTA CRUZ: The picture was taken with our camera [the Wide Field and Planetary Camera], and it looked weird. It was a star, but it had a bright point at the centre. One of the astronomers on our team looked at the image and said, "This telescope has spherical aberration." That immediate diagnosis was extremely severe, with huge consequences.

OLIVIER: The months immediately after launch were just a nightmare.

FABER: Our team wanted to know whether that was really true. We moved the secondary mirror in and out of focus in order to sample the spherical aberration at different levels. In June, at a project meeting, we showed our results and there could be no doubt. It was a catastrophe.

OLIVIER: I got a phone call to come into NASA headquarters. We explained what the problem was. The deputy administrator, J. R. Thompson, kept telling me, "Olivier, you've got to turn another knob on the spacecraft to fix this!" I said, "J. R., I don't have a knob to turn." It took a few days for the top men to realize, deep down in their hearts, that they had a real problem.

We put a telescope in space and it could hardly see. I felt terrible. I felt like a dog wouldn't take a bone from me.

The problem turned out to originate from a spacing error in the device used to shape the primary mirror. The error had been made by the mirror contractor, Perkin-Elmer Corporation, and had been missed repeatedly by NASA. It affected all five of Hubble's initial instruments, and could not be fixed from the ground.

EDWARD WEILER, FORMER HUBBLE CHIEF SCIENTIST: I had the unique honour of being the one to explain what the impacts on the scientific programme of Hubble would be. That was the day of infamy.

But luckily, about two hours before the press conference, [Hubble imaging expert] John Trauger pulled me aside and said: "Ed, I think we've got something you should know about. We think we can fix this. We have these four relay mirrors that are flat, but if we put a small curve on them, a curve that is the opposite of the bad curve on the mirror, it will cancel out."

I reported this to the press conference. I promised we had this fix in hand, and of course nobody believed anything we said. It was not a friendly situation. I had neighbours come up to me and say how much sympathy they had for me working on a national disaster.

FABER: Our big fear that was Hubble would not be fixed. How would we keep the public's and NASA's interest alive in Hubble while a repair plan could be invented?

It took three years to make that plan. NASA engineers had to develop ways to fix each instrument, with all the work done by astronauts in bulky spacesuits working in zero gravity. In December 1993, seven astronauts launched aboard the space shuttle *Endeavour* to save Hubble.

WEILER: If you had asked me for the odds ahead of time, I'd have said 50% success. This was the first time we ever tried to repair a satellite. Five [spacewalks] all had to go perfectly. But things kept going right. It was like a dream sequence. You were afraid you were going to wake up and there was going to be a problem.

We went home at the end of the mission like a surgeon goes home after an eye operation: they've done everything they can, but until the bandages come off you won't know for sure.

ANTONELLA NOTA, ESA HUBBLE PROJECT SCIENTIST, SPACE TELESCOPE SCIENCE INSTITUTE (STSCI), BALTIMORE, MARYLAND: When we saw the first images, it was like history had erased those three years of pain.

WEILER: We were all huddled around a little screen, waiting for the first image to come down. It probably only took five seconds but it seemed like six hours.

First we saw a little dot in the centre, but it was a really well-focused dot. And then we saw the faint stars. You just knew, right then, that we had nailed it. That night, I slept like a baby. The trouble with Hubble was over.

With its corrected vision, the telescope could start doing the science astronomers had always hoped for — including responding to fast-moving celestial events, such as the death of comet Shoemaker–Levy 9, which plunged into Jupiter just months after the repair mission. But that first big test for Hubble was almost a failure.

DAVID LECKRONE, FORMER SENIOR PROJECT SCIENTIST: That was the most exciting week I had on Hubble. Many people don't realize that less than two weeks before the first impact, Hubble went into safe mode. Two days before a critical observation, a software engineer at Goddard [Space Flight Center] figured it out and fixed it. It was a brilliant success, to watch a comet tear apart into fragments and crash into the planet a few months after Hubble had been repaired. Imagine if that had happened in 1993 instead of 1994.

ZOLTAN LEVAY, IMAGE SCIENTIST, STSCI: The first test. That was a huge deal.

**WEILER:** It's a classic great American comeback story.

One after another, Hubble's discoveries began landing on the front pages of newspapers and in top scientific journals.

WEILER: Hubble has been the greatest scientific success in NASA's history. With just one picture it could show how the Universe didn't read our textbooks.

"IT'S A CLASSIC GREAT AMERICAN COMEBACK STORY."

NASA

H. HAMMEL (MIT)/NASA/ESA

Left: John Grunsfeld (right) refurbishes Hubble in 2009. Middle: the impacts on Jupiter of comet Shoemaker–Levy 9. Right: the barred spiral galaxy NGC1300.
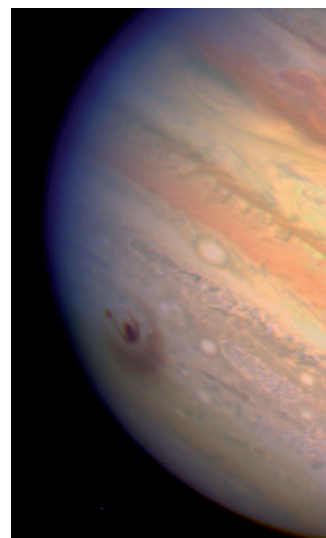
NOTA: Hubble can look in wavelength regimes that are not accessible from the ground, like ultraviolet, because ultraviolet radiation gets absorbed by the atmosphere.

JENNIFER WISEMAN, SENIOR PROJECT SCIENTIST, GODDARD SPACE FLIGHT CENTER, GREENBELT, MARYLAND: There was a burst of new science from Hubble right after 1993. One of these iconic images is the Eagle Nebula, where you see columns of gas where stars have recently formed and are still forming. The informal name is the 'Pillars of Creation', a grandiose title. This gave us a visual clue as to the interaction of young stars.

LECKRONE: Bob O'Dell got pictures of the Orion Nebula. They showed these funny little cocoons all over the place. As you looked more closely, you saw examples of stars surrounded by dark disks. My god, these are places where planets must be forming!

O'DELL: It was the only truly eureka moment I've had as a scientist.

WISEMAN: Hubble homed in on the core of the galaxy M87 to monitor the motion of gas there. The astronomers used a spectrograph to find the gas was moving about a million miles per hour in one direction on one side of the core, and a million miles per hour in the other direction on the opposite side. The only way something could be orbiting this fast would be if there were something very massive in the core in a very small volume. This was the first definitive observation of a supermassive black hole in the core of another galaxy.

LECKRONE: Hubble continues to defy all expectations in creative new ways in which it can be used. Look at dark energy.

KENNETH SEMBACH, HEAD OF THE HUBBLE MISSION OFFICE, STSCI: We know dark energy pervades the Universe because we've been able to measure the expansion rate of the Universe at different times. The key to doing that has been looking at distant supernovae [with Hubble]. The more distant supernovae are dimmer than you would have expected. The teams that won the Nobel Prize in Physics in 2011 realized that the Universe was expanding at an accelerating rate.

This is the equivalent of throwing a ball up in the air and it just decides to speed up and keep going up. That would be a repulsive force rather than an attractive force. It works against gravity.

WISEMAN: The repaired Hubble had exquisite angular resolution that allowed us to look for individual stars, to separate them in crowded regions. In this way you could actually study populations of stars and map out their properties.

The public responded to the flood of gorgeous imagery. Hubble became NASA's first Internet sensation.

LECKRONE: We've developed a following of people who are not astronomers but have learned to love astronomy.

LEVAY: I'm honoured that people admire these results. It has just kind of snowballed. People have done songs and stuff inspired by Hubble. There's poetry, artwork.

We've been batting around ideas of why Hubble is so much in the public consciousness. One is because we came along right when the Internet was really starting to take off. A lot of people had easy instant access to the results from Hubble.

NOTA: We call it the people's telescope. We have really brought the Universe to people's homes. Some 15 years ago I was in this remote area of Papua New Guinea, living on a ship that would dock in places where there wasn't even a harbour. One time, we couldn't believe it, there was a kid wearing a Hubble T-shirt. The child was delighted when we gave him a set of Hubble cards to play with, to go with his T-shirt.

WEILER: After I retired and moved to Florida, I negotiated with my wife. Half the pictures in the house are Hubble, and half are other things.

Astronauts continued to visit the telescope, upgrading and replacing its instruments regularly to extend its life. Sometimes, Hubble's future looked dim. In 1999, astronauts launched an emergency repair mission after three of the telescope's six gyroscopes failed.

JOHN GRUNSFELD, NASA ASTRONOMER AND ASTRONAUT WHO HAS PERFORMED EIGHT SPACEWALKS TO SERVICE HUBBLE: Hubble had gone dark, and it was a real question as to whether the science

"WE CALL IT THE PEOPLE'S TELESCOPE. WE HAVE REALLY BROUGHT THE UNIVERSE TO PEOPLE'S HOMES."

NASA/ESA/HUBBLE HERITAGE TEAM (STSCI/AURA)

↪ NATURE.COM
Scientists select their favourite Hubble images:
go.nature.com/onagtw

"IT WILL BE A GRADUAL, GRACEFUL FAILURE."

A pillar of gas and dust in the Eagle Nebula.

was over. For an astronomer and an astronaut, this was a holy grail of repair missions. Up we went, and soon enough we saw this bright star on the horizon. It was Hubble.

It was surreal. There was one moment when I was out at the end of the robotic arm, and the operator drove me towards Hubble, slowly turning me over. I put out my index finger and just kind of tapped the telescope, to prove to myself it was all real.

We deployed it on Christmas Day. I remember thinking, what better present could there be for planet Earth than a repaired Hubble?

Four years later, in the wake of the Columbia *shuttle disaster, NASA administrator Sean O'Keefe cancelled a final planned servicing mission, citing safety concerns.*

MATT MOUNTAIN, FORMER DIRECTOR, STSCI: What made it worse was the instruments started failing. It was actually pretty bleak. It was clear Hubble was not doing as well as it should be.

WEILER: Luckily administrators changed, and we got Mike Griffin in there. He supported looking at the alternatives, and at the end of the day we got our servicing mission.

MOUNTAIN: Griffin announced he would allocate two shuttles to this. That's an incredible commitment by a space agency to a science mission. Suddenly the attitude changed, and there was a future for the whole team at Hubble.

GRUNSFELD: When we saw it on approach [on the final

servicing mission, in 2009], it was as if we were seeing an old friend. Very few people have hugged Hubble the way I have. I knew all the handrails practically by name. When we let it go, it was in the best shape of its life. We had accomplished our job, and its science heritage would continue.

The telescope remains a premier tool, particularly for time-consuming, data-rich surveys that are meant to benefit the astronomical community for years to come. Hubble set the standard for uploading data to a communal archive available to all astronomers.

JENNIFER LOTZ, ASTRONOMER, STSCI: I feel incredibly lucky to have started my career in the golden age of astronomy and the golden age of Hubble. The idea of saving all the data and making it available to people after a certain amount of time, that was pretty radical. Now it is accepted practice. You don't have to be the student of the most famous professor in the world to have access to the best data in the world.

JASON KALIRAI, ASTRONOMER, STSCI: People have the misconception that its best days are behind it. More than two research papers every day come out of Hubble. What it's doing today is different from what it's done in the past.

NOTA: Look at one example of a topic that didn't even exist when Hubble was launched: exoplanets. When Hubble launched we didn't even know about the existence of planets outside our Solar System. In 25 years that field has completely revolutionized. Hubble was not designed to study exoplanets but now is characterizing their atmospheres. Hubble always surprises us.

NASA is currently testing Hubble's successor, the James Webb Space Telescope, which is scheduled to launch in 2018. But researchers are still planning for Hubble's final years.

WISEMAN: Hubble right now is as scientifically powerful as ever, perhaps more scientifically powerful than ever.

SEMBACH: In the time we have left, we want to push the envelope. We want to do different things that we haven't done before. We've put out a call to the community asking for creative ideas. Should we be devoting more time to specific types of observations? Should we be devoted to helping students do research with the observatory?

We expect to operate through at least 2020. Right now things look pretty good. That gives us a chance to overlap for a year or two with the James Webb Space Telescope.

PAUL HERTZ, DIRECTOR, ASTROPHYSICS DIVISION, NASA: We will operate Hubble as long as it stays scientifically productive. My guess is that something's going to break someday.

LECKRONE: It will be a gradual, graceful failure. With creative engineering you can keep doing good science. As long as we have at least two good instruments, I think we can keep going even when the spacecraft itself has suffered multiple failures. That might take us to 2025. But it's not going to be with us forever, and we're really going to miss it when it's gone. ■ **SEE COMMENT P. 287**

Alexandra Witze writes for Nature *from Boulder, Colorado. Some quotes in this story have been edited for brevity.*

COMMENT
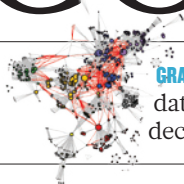
The Hubble Ultra Deep Field 2014 combines the full range of wavelengths available to the telescope, from ultraviolet to near-infrared.

NASA/ESA/TEPLITZ/RAFELSKI/KOEKEMOER/WINDHORST/LEVAY

Hubble's legacy

Twenty-five years after launch, the wild success of the space telescope argues for a new era of bold exploration in the face of tight budgets, says **Mario Livio**.

On 24 April, it will be 25 years since the Hubble Space Telescope (HST) was launched from Cape Canaveral, Florida, into low-Earth orbit aboard the space shuttle *Discovery*.

As well as revolutionizing astrophysics, the first major optical observatory in space — built by NASA with contributions from the European Space Agency (ESA) — has brought the excitement of scientific discovery into millions of homes. Ask people to name a telescope and most will probably say "Hubble".

Circling the Earth every hour and a half, the observatory has completed more than 130,000 orbits and taken more than 1 million exposures of astronomical objects, from dust clouds to distant galaxies. More than 12,800 scientific articles have used HST results, and have been cited more than 550,000 times, making the telescope one of the most productive scientific instruments ever built.

What are the secrets of Hubble's success? Its longevity, pioneering of open data, superior archiving, attention to community needs, dedicated teams of space agencies, astronauts, scientists and engineers, and outstanding outreach infrastructure are all key. These have transformed what seemed initially to be a gigantic failure — flaws in the primary mirror were revealed within weeks — into a scientific triumph.

As Hubble enters its final productive decade, and successors such as the James Webb Space Telescope (JWST) inch towards the launch pad, it is a good time to reflect on its legacy and lessons (see 'Hubble's hits and beyond'). Hubble has taught us that to answer the most intriguing questions in astrophysics, we must think big and put scientific ambition ahead of budgetary concerns. In my view, the next priority should be the search for life beyond our Solar System. A powerful space telescope that can spot biological signatures in the atmospheres of Earth-like exoplanets would be a worthy successor.

ALL IN THE DETAILS

Hubble's greatness lies not so much in the singular discoveries that it has made as in confirming suggestive results from other observatories. As new details have become visible, astrophysicists have had to refine ▶

HUBBLE'S
HITS AND BEYOND

In-flight servicing has prolonged the space telescope's life, paving the way for future missions.

1990
The Hubble Space Telescope is launched on the space shuttle *Discovery* on 24 April. Distortions in the mirror are discovered on 25 June.

1993
In the first servicing mission, astronauts fix the optics and install a new camera.

1996
The first Hubble Deep Field image is released, showing far-flung galaxies.

1997, 1999 & 2002
Servicing missions add a spectrograph and an infrared camera, fix worn gyroscopes that keep the telescope pointing correctly, and replace a camera and solar panels.

2004, 2007
Power supplies fail on the spectrograph (2004) and on a camera (2007).

2008
Hubble shows exoplanet Fomalhaut b and completes its hundred-thousandth orbit of Earth.

2009
Astronauts carry out extensive repairs, and install a new camera and spectrograph.

2011
Hubble makes its millionth observation (of an exoplanet) and the ten-thousandth scientific paper using its data is published (concerning supernovae).

2018
The James Webb Space Telescope will open up infrared views of the Universe.

~ 2024
WFIRST/AFTA will enable large surveys in the infrared from space.

~ 2030
Proposed launch of a new major observatory to image and characterize exoplanets.

their theories about the Universe.

The telescope's power stems from its high perch above most of Earth's atmosphere, at an altitude of about 560 kilometres. Unaffected by airglow (faint light emitted by atmospheric chemical processes) and turbulence, Hubble has a sharp eye (resolution) and can detect faint objects (sensitivity), even though its 2.4-metre-diameter mirror is small by today's standards (8–10-metre mirrors are now the norm). It can resolve objects 0.07 arcseconds apart — akin to reading the year on a dime from three kilometres away. That is ten times finer in visible light than any ground-based observatory can achieve.

The telescope sees wavelengths from ultraviolet to the near-infrared, including bands that are blocked by the atmosphere to astronomers on Earth. Its capabilities in the ultraviolet are about 100 times greater than those of its predecessors or of any current telescope.

The original plan for Hubble was for it to tackle three major problems: measure how fast the Universe is expanding, work out how galaxies evolve, and probe the structure of diffuse gas clouds lying between galaxies (the intergalactic medium). It has succeeded and provided unexpected sights along the way. Here is my selection of a few of Hubble's most important scientific achievements.

GREATEST HITS

One of the telescope's first jobs was to reduce the uncertainty in the cosmic expansion rate — the 'Hubble constant' — named, like the telescope, after its discoverer Edwin Hubble. Between 1994 and 2011, the uncertainty was reduced from a factor of 2 to a few per cent. Hubble thus helped to set the age of the Universe at 13.8 billion years. It did so by extending to more remote galaxies an established method of inferring distances from the cycles of changing brightness in a class of pulsating stars known as Cepheid variables.

The HST confirmed in 1998 that the cosmic expansion is accelerating, propelled by a mysterious form of 'dark energy'. This feat was achieved by monitoring supernovae — exploding stars — that are out of the reach of ground-based telescopes. Understanding the nature of dark energy is one of the most important challenges that physicists face.

The telescope also produced an 'executive summary' of star formation across cosmic time. In a series of roughly tenday observations between 1995 and 2014, it peered intently at small patches of sky, reaching deeper than any instrument has gone before. The resulting images are collectively known as the Hubble Deep Fields. Finding that many galaxies already existed 500 million years after the Big Bang, the HST challenged ideas about how the first stars formed, heated and re-ionized the Universe. Astronomers are still trying to fully understand why the rate at which new stars were

born peaked about 10 billion years ago.

Using its high resolution to observe the motions of stars and gas in the centres of galaxies, the Hubble telescope proved that almost all galaxies have at their heart a supermassive black hole (with masses millions to billions of times that of the Sun). The mass of the black hole scales with that of the 'bulge' of stars surrounding it, showing that galaxies and black holes evolved together.

The HST also determined for the first time the chemical composition of the atmospheres of some giant extrasolar planets, revealing in 2001 the spectral signatures of elements such as sodium and in 2008 molecules such as water and methane. A larger telescope might one day be able to identify signatures of life processes — such as oxygen and chlorophyll — in the atmospheres of rocky planets beyond our Solar System.

SECRETS OF SUCCESS

Scientific prowess is not the sole reason for Hubble's success. Five servicing missions — in 1993, 1997, 1999, 2002 and 2009 — by space-shuttle astronauts allowed the telescope to be reinvented. Astronauts have introduced corrective optics, replaced mechanical tape recorders with solid-state memory drives, upgraded the solar arrays and installed cameras and spectrographs. Without those repairs, Hubble would not be working today, or would be operating with 1970s technology.

Chance favours the prepared. Four more factors have multiplied the HST's productivity: making data rapidly and openly available; effective and accessible archiving; undertaking risky projects; and a robust funding and fellowship system.

Creative thinking was championed through reserving 10% of observing time for very large, time-critical or unconventional proposals at the director's discretion. The original Hubble Deep Field imaging, for instance, was advocated and led by Robert Williams, then director of Hubble's scientific operator, the Space Telescope Science Institute (STScI). Other observatories, including the Gemini Observatory in Hawaii and Chile and the Large Binocular Telescope in Arizona, have adopted the approach.

Researchers are given a year to analyse Hubble observations before the data are made public. Special data sets such as the Hubble Deep Fields were made openly accessible immediately. The HST was not the first space observatory to adopt this policy, but it inspired others to follow suit: the data from the Swift Gamma-Ray Burst Mission, launched in 2004, for example, are immediately available.

From the start, the archiving and dissemination of data were more rigorous and more highly automated (including calibrated data, for example) than at other observatories. For the past decade, more archive-based papers

have been published each year than ones using proprietary data: in 2014, 302 papers relied on archival data alone; 283 used proprietary data. The European Southern Observatory adopted the HST archiving practices in 1993.

All allocated HST observations come with a NASA research grant, to ensure that the data are analysed and the results published quickly. Since 1990, more than 4,600 HST proposals have been accepted, and grants awarded totalling US$500 million.

The project has also sponsored a new generation of top researchers. Since 1990, there have been 352 Hubble fellows — postdoctoral researchers who are funded to work independently for three years on Hubble-related science at US universities. Since 1993, about 500 PhD theses have used Hubble data.

THE PEOPLE'S TELESCOPE

The HST has transformed the landscape of scientific outreach and education. An STScI Office of Public Outreach was funded almost from the start to offer press releases, online outreach and education to schools, science centres and planetariums. Embedding the office in the STScI — located on the campus of Johns Hopkins University in Baltimore, Maryland — ensured that professional astronomers were involved. An attractive and user-friendly website (hubblesite.org) attracts billions of hits a year.

Hubble educators pioneered the online dissemination of materials to schools, starting at a time when little was available. Today its materials reach more than 6 million students and 500,000 educators each year in the United States alone. Multimedia presentations on galaxies, exoplanets and black holes play in science centres worldwide.

Hubble images — dubbed by British art critic Jonathan Jones "the most flamboyantly beautiful artworks of our time" — have infiltrated general culture. A dedicated team ensures their visual quality. HST views have been included in art exhibits from Baltimore to Venice. They adorn book covers and music albums, such as *Binaural* by the rock band Pearl Jam, have inspired contemporary classical music (such as *The Hubble Cantata* by composer Paola Prestini) and dance performances.

> *"For the first time in human history, an answer to the question 'Are we alone?' is within reach."*

NOW WHAT?

The Hubble has shown that it is better to fund the right experiment fully than to compromise to fit a tight budget. Likewise, future major astronomical endeavours should: identify the most important question that needs to be answered; determine what it would

The Tarantula nebula, snapped by Hubble in visible, infrared and ultraviolet light.

take to answer the question and the technical feasibility of doing so; estimate the full cost of such a project; evaluate whether the goal is worth the investment; and act accordingly. That is, establish the necessary funding profile and keep it stable. Avoid cost overruns through careful planning and oversight.

The most intriguing question in astronomy is, in my view, whether life exists in our Galaxy beyond the Solar System. Thanks especially to the Kepler space telescope, we know that the Galaxy is teeming with hundreds of millions of Earth-sized planets in the 'habitable zones' of their host stars that allow for liquid water on a rocky surface.

The next steps are laid out. When it is launched in 2017, the Transiting Exoplanet Survey Satellite (TESS) should find a handful of nearby planets slightly heavier than Earth in the habitable zones of low-mass stars. The orbital periods of such planets are short and their stars faint, making them somewhat easier to detect. Then, the JWST, to be launched in 2018, and the Wide Field Infrared Survey Telescope–Astrophysics Focused Telescope Assets (WFIRST/AFTA), planned for around 2024, should look for water and other molecules in the atmospheres of a few of these planets.

A more powerful telescope will be needed to place meaningful statistical constraints on how common or rare life in the Galaxy is. One with a mirror at least 12 metres across and with a resolution 25 times that of Hubble's would be able to image a planet next to its star and detect spectrally the presence of oxygen and other biosignatures in its atmosphere. WFIRST/AFTA should be able to detect a planet 1 billion times fainter than its star; a brightness contrast of 10 billion will be required to image an Earth analogue next to a Sun-like host star. Clearly, such a telescope would offer a plethora of other discoveries as well.

A large sample of planets — around 50 — would have to be tested. Calculations show, for example, that if no biosignatures are detected in more than about three dozen Earth analogues, the probability of remotely detectable extrasolar life in our Galactic neighbourhood is less than about 10%.

A report on such a 'high-definition' telescope is expected to be published around June by the Association of Universities for Research in Astronomy. Several steps should be taken now. First, NASA, ESA and other potential international partners should convene a panel to examine such a project. Technology-development studies should be accelerated to make a launch around 2030 plausible. The search for life must be prioritized in the next US and international decadal surveys that guide national funding decisions about missions. The US astronomical community will recommence those discussions in 2016 for research priorities in the next decade.

In the meantime, I would also welcome substantially increased investment in the Search for Extraterrestrial Intelligence (SETI) project. Around $100 million in extra funding, perhaps from private sources, would speed up the survey to a point at which about 10 million stars could be searched in a decade for radio or optical signals that are indicative of intelligent life. The chance of success may be low, but the pay-off could be huge.

For the first time in human history, an answer to the question 'Are we alone?' is within reach. The search for life should be high on the scientific agenda for the next 25 years. ■ **SEE NEWS FEATURE P.282**

Mario Livio *is an astrophysicist at the Space Telescope Science Institute (STScI) in Baltimore, Maryland, USA.*
e-mail: mlivio@stsci.edu

Make precision medicine work for cancer care

To get targeted treatments to more cancer patients pair genomic data with clinical data, and make the information widely accessible, urges **Mark A. Rubin**.

Ten months ago, the physicians of a feisty 76-year-old sales clerk from New Jersey who had an advanced carcinoma in her urinary tract decided to try an unconventional therapy. A few weeks earlier, they had sent a sample of her tumour to my team at the Institute of Precision Medicine at Weill Cornell Medical College and NewYork-Presbyterian Hospital in New York City. Genetic sequencing had revealed that she had more copies than usual of the *HER2* gene (also known as *ERBB2*)[1,2].

After years of failure with the usual arsenal of surgery, chemotherapy and radiation, the physicians included the drug Herceptin (trastuzumab) in the woman's treatment. Herceptin is more commonly used for breast cancer, but it targets the *HER2* mutation. Since taking the drug, she has been free of disease.

Advances in sequencing have dramatically increased the likelihood of discovering mutations that drive tumour growth in certain people and in certain tumours — even in specific cells within tumours. Yet mountains of genomic data are accumulating that are of little use because they are not tied to clinical information, such as family medical history. What is more, genomic data are generally confined to documents that cannot easily be searched, shared or even understood by most physicians.

To achieve the level of success in precision medicine for cancer care that US President Barack Obama and others are anticipating,

sequence data needs to be linked, in real time, to the patient sitting in front of his or her doctor. Integrated genomic and clinical data will also need to be available, in a searchable way, to a broad community of practitioners and researchers. Prototypes for centralized data banks are showing promise, but serious and sustained investment is needed to scale them up.

COMPLEX RECORDS

Clinicians are used to appraising 20–50 measurements from routine laboratory tests, such as for blood-sugar levels. Such data can be easily entered into patients' electronic health records. Genomic data introduces a whole new level of complexity.

To give an idea of the scale, it would take more than 25 days to transfer from one computer server to another the 2.5 petabytes (a petabyte is 1,000 terabytes) of data generated by The Cancer Genome Atlas — a US project started in 2005 to catalogue the mutations that drive cancer. This is according to my colleague Toby Bloom, deputy director for informatics at the New York Genome Center, a consortium that specializes in large-scale human genome sequencing.

Hugely complicated genomic reports are rarely available in electronic form and are seldom tied to basic information about the patient. Whole-genome sequencing on tumour samples from nearly 14,000 people by the International Cancer Genome Consortium (ICGC), for instance, has revealed nearly 13 million mutations across the genome. But numerous factors aside from the mutations in a person's DNA will affect whether any one patient will respond to a particular treatment. Unfortunately, in the ICGC effort — and many like it — only the most minimal of clinical data, such as type and size of a tumour, are available (see 'Missing metrics').

Since 2013, working with a team of computational biologists from Weill Cornell and the Centre for Integrative Biology at the University of Trento in Italy, my colleagues and I have conducted a pilot programme to determine the feasibility of tying genomic to clinical data in real time. So far, we have created easy-to-read reports for 250 people with cancer.

Each report carries a barcode, allowing patients to be de-identified and re-identified as needed, and is designed to be integrated easily into the electronic health-records system of the NewYork-Presbyterian Weill Cornell Medical Center. The data, which are presented much like pathology results, capture clinical information (family history, medication use and so on), information about mutations for which specific drugs exist, and findings about genetic anomalies with unknown effects.

We have discovered that more than 90% of our patients carry a mutation that may be responsive to a known drug — although less than 10% of the patients may be eligible for a clinical trial either for logistical reasons or because there is insufficient evidence to warrant trying a non-approved drug.

To be useful more broadly, these data need to be sharable across institutions. Take, for instance, current efforts to investigate the efficacy and safety of the drug neratinib in patients whose tumour growth is driven by various mutations in either *HER2* or *EGFR*[3]. Aside from lung cancer (in which *EGFR* mutations are common), the frequency of these mutations is in the range of 1–6%, so achieving the numbers required for a phase II clinical trial has meant recruiting patients from multiple medical centres. Sharing data across institutions could dramatically increase the ease and efficiency of recruitment for such trials — currently a frustratingly slow process that is largely dependent on word of mouth.

Yet the barriers to achieving this type of sharing are formidable. In the United States, incompatible electronic systems make transferring patient records between facilities extremely difficult — often requiring the shipping and scanning of printouts.

DIGITAL DATA
Various initiatives are trying to address the creation of standards for communal digital medical data. One example is the non-profit New York City Clinical Data Research Network (NYC-CDRN). Funded by the Patient-Centered Outcomes Research Institute in Washington DC, this non-governmental organization is bringing together 22 institutions, led by the Weill Cornell Medical College and NewYork-Presbyterian Hospital, to document and manage clinical data[4].

Sixteen months in, the NYC-CDRN has more than 6 million records with hundreds of thousands of data elements, ranging from simple measurements of, say, calcium levels in the blood, to the results of magnetic-resonance-imaging scans. The ultimate goal is to include genomic data in the database and to follow patients longitudinally. Particularly in countries with private health-care systems, centralized 'warehouses' of shared, standardized, searchable patient data may be the most feasible way forward.

The promise of precision medicine for cancer is now clearly evident. For instance, drugs that target *BRAF(V600E)* mutations (seen in around 60% of melanomas) and *IDH1* or *IDH2* mutations (seen in around 80% of brain tumours) have either been approved or are undergoing testing in clinical trials[5,6] — although, as with most targeted therapies, resistance is a major problem[7]. And in one of the most ambitious precision-medicine trials ever conducted, which is taking place at multiple institutions in France, 141 patients out of the 708 enrolled have already been matched to targeted-therapy trials[8].

MONEY MATTERS
Yet the 'precision' approach raises some hard questions. The more patient-specific information included in centralized databases — crucial to the long-term success of precision medicine — the harder it will be to ensure contributors' anonymity. What rights should people have over their own health data? Should such data be shared internationally? Also unclear is who should manage and sustain such data warehouses, and who should pay for them.

The NYC-CDRN has already cost US$7 million, and annual costs will increase as more information is collated. This adds to the considerable expense of the treatments themselves — annual costs for the targeted therapies in cancer now available generally exceed $100,000, and most extend patients' lives by only months.

Should targeted drugs for patients with mutations found in only 10% of the population be developed and used if they extend survival by just three months, say? Should drugs be supported only if they extend people's lives for at least one year?

To complicate things, the full benefits of many drugs may become apparent only after they have been approved. Herceptin, for instance, was initially approved by the US Food and Drug Administration as a treatment that can extend the survival of people with a certain advanced metastatic breast cancer by months[9]. Increased use of the drug has since revealed that it can improve the chances of long-term survival for people with earlier stages of breast cancer[10].

"Incompatible electronic systems make transferring patient records between facilities extremely difficult"

Some organizations have already given guidance on the rationing of precision treatments. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) examined data on the usefulness of different types of genomic test in the treatment of breast cancer. In September 2013, NICE recommended a test called Oncotype DX for clinical decision making but determined that three other genomic tests currently available (MammaPrint, IHC4 and Mammostrat) be used only in research because of insufficient evidence supporting their usefulness in clinical care.
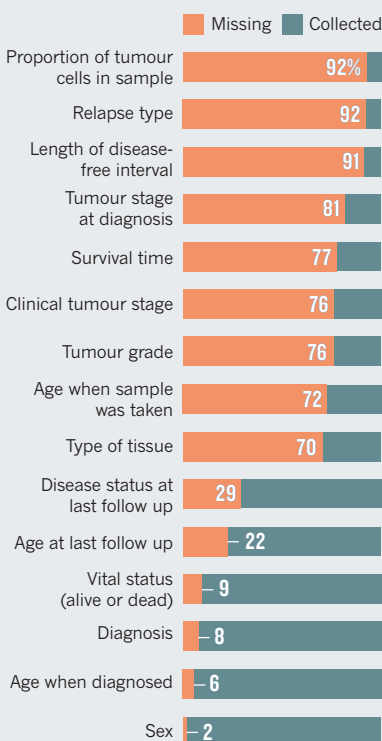
There are many reasons for hope. But turning the wealth of insights potentially available from genomics into targeted treatments for cancer will require difficult decisions and the costly, laborious task of creating shared and searchable information. ∎

Mark A. Rubin *is professor of oncology in pathology and director of the Institute for Precision Medicine at Weill Cornell Medical College and NewYork-Presbyterian Hospital in New York City, New York, USA.*
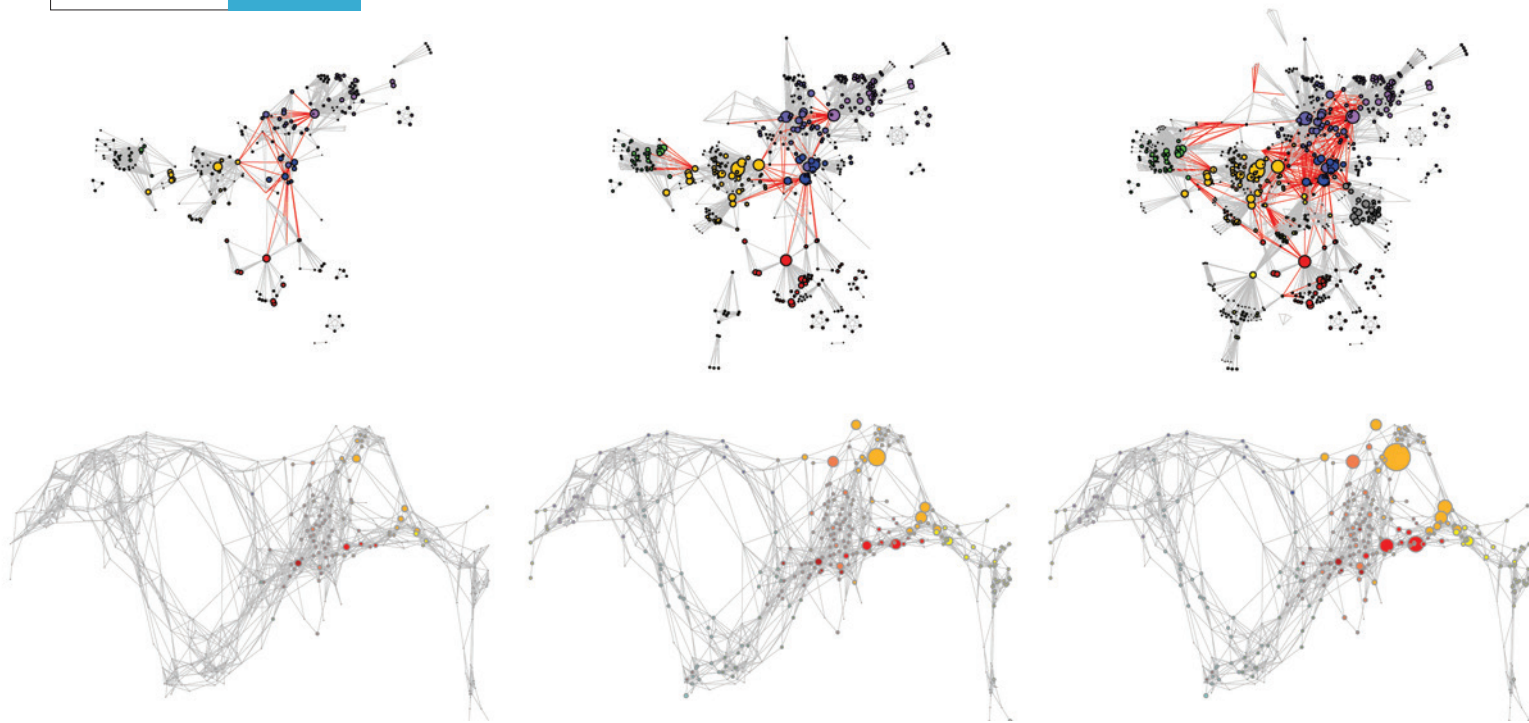*e-mail: rubinma@med.cornell.edu*

1. King, C. R., Kraus, M. H. & Aaronson, S. A. *Science* **229,** 974–976 (1985).
2. Hudziak, R. M., Schlessinger, J. & Ullrich, A. *Proc. Natl Acad. Sci. USA* **84,** 7159–7163 (1987).
3. Rabindran, S. K. *et al. Cancer Res.* **64,** 3958–3965 (2004).
4. Kaushal, R. *et al. J. Am. Med. Informat. Assoc.* **21,** 587–590 (2014).
5. Flaherty, K. T. *et al. N. Eng. J. Med.* **363,** 809–819 (2010).
6. Wang, F. *et al. Science* **340,** 622–626 (2013).
7. Ramos, P. & Bentires-Alj, M. *Oncogene* http://dx.doi.org/10.1038/onc.2014.314 (2014).
8. Massard, C. *Ann. Oncol.* **26,** (suppl. 2) ii4 (2015).
9. Slamon, D. J. *et al. N. Eng. J. Med.* **344,** 783–792 (2001).
10. Slamon, D. *et al. N. Eng. J. Med.* **365,** 1273–1283 (2011).

MISSING METRICS
For much of the genomic data obtained for nearly 14,000 patients by the International Cancer Genome Consortium, key clinical information is missing.

Missing / Collected

Metric	Missing (%)
Proportion of tumour cells in sample	92%
Relapse type	92
Length of disease-free interval	91
Tumour stage at diagnosis	81
Survival time	77
Clinical tumour stage	76
Tumour grade	76
Age when sample was taken	72
Type of tissue	70
Disease status at last follow up	29
Age at last follow up	22
Vital status (alive or dead)	9
Diagnosis	8
Age when diagnosed	6
Sex	2

A visualization maps the impact of US National Institutes of Health funding strategies on authorship networks (top) and publication output (bottom).

DATA VISUALIZATION

Mapping the topical space

Rikke Schmidt Kjærgaard applauds a cogent guide to scientific cartography.

Flip through ten pages of this issue of *Nature*, and your eyes will be drawn to headlines and images. In our information-thick, data-supported world, optimal representation is key. Yet many scientists lack the tools and training to create great data visualization — to digitally parse data in many dimensions, revealing patterns and relationships in phenomena ranging from patent citations to the evolution of great scientific discoveries.

Guidance is on offer from books such as Edward Tufte's *Envisioning Information* (Graphics Press, 1990) and Stephen Few's *Show Me the Numbers* (Analytics Press, 2004). In recent years, *Nature Methods'* Points of View column by Bang Wong, Martin Kryzwinski and invited co-authors has tested design rules on real data sets (see go.nature.com/3scjfr). Now, in *Atlas of Knowledge*, information scientist Katy Börner aims to bring much of this together. As the second book in a series of three, it follows *Atlas of Science* (MIT Press, 2010), an introduction to the power of information visualization (see B. Schneiderman *Nature* **468**, 1037; 2010). Both books complement Börner's comprehensive travelling exhibition *Places & Spaces: Mapping Science*, now in its tenth year (http://scimaps.org).

In *Atlas of Knowledge*, Börner gives guidance on how to 'map' — make visualizations of statistical, temporal, geospatial, topical and network data to aid intelligent decision-making by scientists, economists and policy-makers. One standout example is the beautiful 2011 'Design vs Emergence: Visualization of Knowledge Orders' by Alkim Almila Akdag Salah and her colleagues, which compares Wikipedia's category structure with the Universal Decimal Classification system. The book as a whole is an impressive, visually captivating resource, although ultimately it is more a tour inviting comparison and inspiration than a step-by-step manual.

In part 1, Börner first explores research at the micro level, such as the evaluation of individual scholarly merit on the basis of citation counts, prestige, internationalization and funding. She progresses by stages to multilevel and universal research, including investigating population size, life expectancy, national debts and gross domestic product on a global scale. Part 2 introduces valuable techniques for general data analysis and visualization, including how to

Atlas of Knowledge: Anyone Can Map
KATY BÖRNER
MIT Press: 2015.

map geospatial location, correlations and relationships, trends and distribution. Börner presents an encyclopaedia of examples of needs-driven workflow design and data scale, as well as types of visualization such as tables, charts, graphs, maps and networks.

The practical value of the book lies in bringing these case studies together to evaluate the pros and cons of different strategies in visualization design. The variety is breathtaking. An example of Hans Rosling's Gapminder visualizations, for instance, lays out global socio-economic data for 1930–2012; derived from Rosling's graph Wealth & Health of Nations, it was crafted with the Trendalyzer software that he developed for animating statistics. And Ben Fry's 'On the Origin of Species: The Preservation of Favoured Traces' (http://benfry.com/traces/) compares editions of Darwin's magnum opus using Processing, an open-source programming language used to teach computational design. Both pack comprehensive data into easy-to-read graphics, utilizing variables such as colour, geometry, statistics and development over time.

Part 3 is where *Atlas of Knowledge* stands out from other treatments, presenting 40 full-page iconic images authored by pioneers of data visualization. The visionary US architect Buckminster Fuller, for instance, was — with artist and sociologist
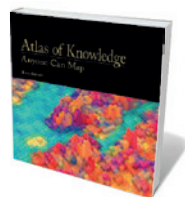
HANS ROSLING

Books in brief

John McHale — one of the first to chart long-term trends of industrialization and globalization. The 1965 chart 'Shrinking of Our Planet by Man's Increased Travel and Communication Speeds Around the Globe' maps how the confluence of communication and transportation technologies from 500,000 BC to 1965 have conquered distance.

However, you will need to go to the *Places & Spaces* website to fully appreciate the complexity and interactivity of many of the twenty-first-century digital visualizations. For example, in print it is hard to locate the bacterium *Streptococcus pneumoniae* on the 2006 'Tree of Life' map by Peer Bork and his colleagues, which shows 191 species with fully sequenced genomes. Moreover, the wealth of examples and illustrations in Börner's book is sometimes a bit too rich. With fewer images, it would have been possible to lead readers into the details, allowing us to see what is at stake without running back and forth between book and screen.

Atlas of Knowledge places itself in a long line of resources on data visualization. The focus is less on how-to than it was in, say, Felice Frankel and Angela DePace's *Visual Strategies* (Yale Univ. Press, 2012), but Börner's book has a place on my shelf. Whether you read it cover to cover or just browse the extraordinary examples, you put it down inspired. ∎

Rikke Schmidt Kjærgaard *is associate professor of scientific data visualization and head of the Visualization Lab at Aarhus Institute of Advanced Studies, Aarhus University, Denmark.*
e-mail: risk@aias.au.dk

Beyond: Our Future in Space
Chris Impey W. W. NORTON *(2015)*
Does navigating a pure vacuum while "strapped to a barely controlled chemical explosion" appeal? Yes — to a select proportion of us, notes astronomer Chris Impey in this bold, elegant and engaging exploration of space travel past, present and future. Impey ranges widely, over a variant of the dopamine-controlling gene *DRD4* that may encourage astronauts to seek novelty; the work of visionaries such as rocket scientist Konstantin Tsiolkovsky; the trajectories of national space programmes; advances in robotics and exoplanet discoveries; the potential for extraterrestrial life; and far beyond.

The Prime of Life: A History of Modern Adulthood
Steven Mintz BELKNAP *(2015)*
Coming of age, argues historian Steven Mintz, is not what it used to be. Characterizing adulthood as a "historical black hole", Mintz sets out to trace the concept's trajectory from the nineteenth century to its 1950s apex, and its disintegration in our individualistic times. He looks at shifts in intimacy, marriage, parenthood and work, noting that some 80% of today's US citizens in their late twenties have yet to tick off all the traditional indicators of adulthood, such as leaving home. Yet we need to dig deeper to redefine adulthood, he avers — not least, by reinstating qualities such as judgement to the definition.
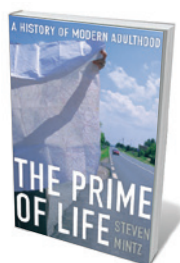
Einstein's Dice and Schrödinger's Cat: How Two Great Minds Battled Quantum Randomness to Create a Unified Theory of Physics
Paul Halpern BASIC *(2015)*
Physicist Paul Halpern tells the entangled tale of Albert Einstein, Erwin Schrödinger and their search for a Grand Unified Theory with humour and concision. Schrödinger allied himself with Einstein to counter the orthodox quantum view championed by Niels Bohr and others. But as Halpern reminds, Schrödinger was as contradictory as his famous thought experiment, and Einstein was prone to premature announcements of theoretical success. A spat between them, he shows, deprived them of further collaboration, and us of the fruits.
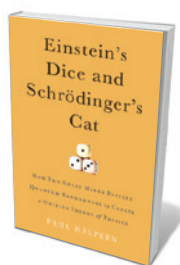
The World Beyond Your Head: On Becoming an Individual in an Age of Distraction
Matthew B. Crawford FARRAR, STRAUS AND GIROUX *(2015)*
In this follow-up to his *Shop Class as Soulcraft* (Penguin, 2009), philosopher-mechanic Matthew Crawford looks at the toll that the assault of constant advertisements, mobile-phone calls and more are having on our collective psyche. The resulting fragmentation and dissociation are well documented. Crawford's solutions — creating an "ethics of attention" and reclaiming "the real" through, for instance, craft — are pragmatic, but the rather belaboured philosophical overlay sometimes wars with his message.
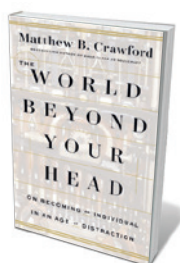
The Archaeology of Sanitation in Roman Italy: Toilets, Sewers, and Water Systems
Ann Olga Koloski-Ostrow UNIV. NORTH CAROLINA PRESS *(2015)*
From aqueducts to amphitheatres, ancient Rome was a hotbed of engineering. That ingenuity percolated downwards too, as classicist Ann Olga Koloski-Ostrow shows in this uneven yet enlightening treatise on sanitation in Roman Italy in the first centuries BC and AD. Homing in on Herculaneum, Ostia, Pompeii and Rome, she explores sanitation design, concepts of hygiene and the role of scatology in the literature and public-toilet graffiti of the time. Barbara Kiser

Q&A John Henry

Geological historian

The first geological map of a nation was made 200 years ago by British surveyor William Smith; the rediscovery of a first-edition copy in the archives of the Geological Society of London was announced last month (see go.nature.com/oogpht). As researchers gather for a conference to celebrate the anniversary of the 1815 chart of England and Wales, John Henry, chair of the society's history group, talks about the map and its pioneering creator.

How did William Smith get his start?
He was a very practical man with a weak formal education — he left school when he was 11 years old. You would not have expected him, a boy from a small village in Oxfordshire, to have accomplished what he did. His ability to draw and to observe spurred his uncle to get him books on geology and surveying. Eventually Smith was apprenticed to a surveyor, and he was off on his first assignment at just 22.

How did he develop his geological ideas?
He began to work in Somerset, surveying the routes for a canal to carry coal to market. Going down mine shafts to study the thickness of core seams and the distances between them, he noticed the different layers of rocks and the fossils they held. He saw that fossils that looked broadly similar were actually slightly different, depending on which rock strata they were in. In his memoirs, he recalled that he was trying to make a three-dimensional model of the landscape. He worked out that particular fossils are found in particular rocks, and that the rocks are always in the same sequence. Nobody else had picked up on that.

William 'Strata' Smith's 1815 map charted the rocks around part of Britain.

Why was this significant?
There was no precedent for his concept of geology. Before Smith, people mapped rock by layers and not by fossil content. Smith's approach showed him where he was. Was he below coal? Was he above coal? There were times when he said to landowners not to waste their time drilling in a certain place, because there was no coal. This had a huge economic impact, and remains the fundamental concept underpinning modern prospecting, and the oil industry in particular. Scientifically, Smith's work formed the basis for everyone who came after him. He is the father of English geology.

William Smith Meeting 2015: 200 Years of Smith's Map
23–24 April. Burlington House, London.

What was the wider context for Smith's work?
In the early nineteenth century, there was no systematic mapping of the whole country. Smith was carrying so much in his head, and fleshing it out as he travelled. In his busiest period of consultancy, he covered perhaps 16,000 kilometres, on horseback, walking and in carriages.

How was his map received by colleagues?
The Geological Society started out as a gentlemen's club. Smith was not part of that; he was rural working class. But the aristocrats who employed him could see that he got results in terms of draining land, stabilizing slopes and holding back the sea. So he had a lot of powerful friends, including the naturalist and Royal Society president Joseph Banks, who supported him.

The Geological Society's 1820 geological map resembles Smith's. Did it plagiarize his work?
It muddies the water. It was a team effort coordinated and compiled by geologist George Greenough. He certainly got a head start by having a look at Smith's map, and that was always Greenough's argument: it is the same underlying geology we're mapping, so of course it looks the same. No one really believed him, yet it was not until 1865 that the society took Greenough's name off that map and acknowledged Smith as a major source.

What happened to Smith after his map was published?
He had managed his finances badly and had a financial failure almost as soon as his map came out in 1815. He had to sell his fossil collection and let go of his London and Somerset properties, and he briefly spent time in debtor's prison. But there was a turnaround by 1831. The Geological Society gave him its Wollaston Medal. Smith was pleased to be recognized, and his fortunes began to recover. He spent the last two decades of his life in Yorkshire, and it was a very sunny period for him: clearly he had become a grand old man of geology. His nephew John Phillips went on to be a great geologist, a professor of geology at the University of Oxford and a driving force behind the Oxford University Museum of Natural History.

Why is Smith's map still so important today?
It is all part of building blocks. Smith got the concept right, and other people came along and built on that. Later mappers were able to carry the concept through to more difficult terrains with a more complex history of folding and faulting, as in Scotland. But everything starts with him. ∎

INTERVIEW BY ALEXANDRA WITZE

Correspondence

Is a mega-project the ELI in the room?

Romania is firmly committed to contributing to the prestigious Extreme Light Infrastructure (ELI) nuclear physics project, co-funded by the European Regional Development Fund (www.eli-np.ro). As a director of university research grants — hence one of the many possible competitors of ELI for government funding — I support this commitment.

Started in 2013 and due to go live in 2018, ELI is already competing in budget size with the more than US$100 million that represents the annual total offered by the country's open-competition national grant schemes (see go.nature.com/raad8w).

Romania's national research budget has been notably stable over the past five years — and yet individual grants have been shrinking. Explanations have included the relative priorities of different fields or grant types, the international financial crisis, and ethical issues (see, for example, go.nature.com/j8slvh).

With the economic problems fading away and ELI advancing rapidly, an increase in the national research budget would seem logical — and would in fact have precedents. Alternatively, Romania's research ministry may wish to seek support from other ministries, such as those that specialize in infrastructure.

Radu Silaghi-Dumitrescu *Babeş-Bolyai University, Cluj-Napoca, Romania.*
rsilaghi@chem.ubbcluj.ro

Antibodies: validate recombinants once

It goes without saying that recombinant antibodies, like all binding reagents, need to be validated at the outset (see R. D. Polakiewicz *Nature* **518,** 483 (2015) and L. P. Freedman *Nature* **518,** 483 (2015)). However, we anticipate that recombinant antibodies will require only one such extensive characterization — unlike conventionally raised antibodies.

This single validation will assure scientists that antibodies with identical sequences will have similar reactivity profiles — subject to routine checks that binding activity has not been compromised during transit or by storage conditions.

We are aware that our proposal is incompatible with current business models for commercial reagent antibodies. We do not believe that the answer is to defend the status quo, which has not served science well (A. Bradbury and A. Plückthun *Nature* **518,** 27–29; 2015). The solution is to develop more imaginative business strategies that are compatible with the marketing of fully validated, publicly available recombinant antibody sequences.

Andrew M. Bradbury *Los Alamos National Laboratory, New Mexico, USA.*
Andreas Plückthun *University of Zurich, Switzerland.*
amb@lanl.gov

Inform public on GM, don't cheerlead

Qiang Wang urges China's scientists to support the government in convincing a sceptical public about the benefits of genetic modification (GM) of agricultural crops (*Nature* **519,** 7; 2015). But there is a distinct line between improving scientific communication and cheerleading for the technology itself.

Scientists should not be in the business of "persuading" the public, nor should they compromise their credibility through hyperbole and oversimplification. Their role is to collect data objectively and use the information to accurately convey the possible risks and benefits.

Scientists should never feel compelled to take sides in polemics, only to present the facts as they understand them.

Yongbo Liu, Junsheng Li *Chinese Research Academy of Environmental Sciences, Beijing, China.*
C. Neal Stewart Jr *University of Tennessee, Knoxville, Tennessee, USA.*
liuyb@craes.org.cn

Zero net emissions from Venter facility

You raise the difficult question of reducing the huge carbon footprint associated with research institutions (*Nature* **519,** 261; 2015). We draw your attention to the J. Craig Venter Institute in California: a clean, green scientific research building that could be a model for others.

Decisions made on building designs now will affect carbon emissions for many decades. In designing the new institute, a genomics research facility, we took responsibility for drastically cutting carbon emissions from its daily operations for the next 50 years or so.

We completely covered the roof with photovoltaic panels, which generate 485 kilowatts of power, which alone would have met only 25% of the energy needs of a typical 4,200-square-metre building. Yet the innovative building design is highly energy-efficient and cuts energy demand by 75% through, for example, heating and cooling with water rather than air; recovery and reuse of 'waste' heat from the water-cooled, −80 °C freezers; chemical sensors that allow fewer air exchanges each hour while improving lab safety; deploying operable windows in the office wing; and using natural daylight throughout.

As a result, we have created a laboratory workplace that operates with zero net carbon emissions.

J. Craig Venter, Robert M. Friedman *J. Craig Venter Institute, La Jolla, California, USA.*
rfriedman@jcvi.org

Bury botany's outdated image

Botany courses at academic institutions are dwindling worldwide, yielding too few graduates to replace retiring botanists (see, for example, go.nature.com/sdcagw). Yet botanical expertise is fundamental to a range of topical issues, including biodiversity, agricultural development, biofuel production, drug discovery and food science.

The public can find it hard to differentiate between even common plants, so botanists should engage more in outreach efforts. They also need to devise fresh approaches to teaching upcoming generations about the importance of plants, relying less on pressed dead specimens and focusing on new molecular and systematic tools. This more contemporary treatment of the subject could help to counter botany's lack of appeal to students and research-funding agencies.

Only then can courses move beyond standard taxonomy to important applications such as the discovery of new genes and gene functions. Let public outreach bury botany's old-fashioned image once and for all.

Isabel Marques *University of British Columbia, Vancouver, Canada; and University of Zaragoza, Huesca, Spain.*
isabel.ic@gmail.com

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com** after consulting the guidelines at **http://go.nature.com/cmchno**. Alternatively, readers may comment online: **www.nature.com/nature**.
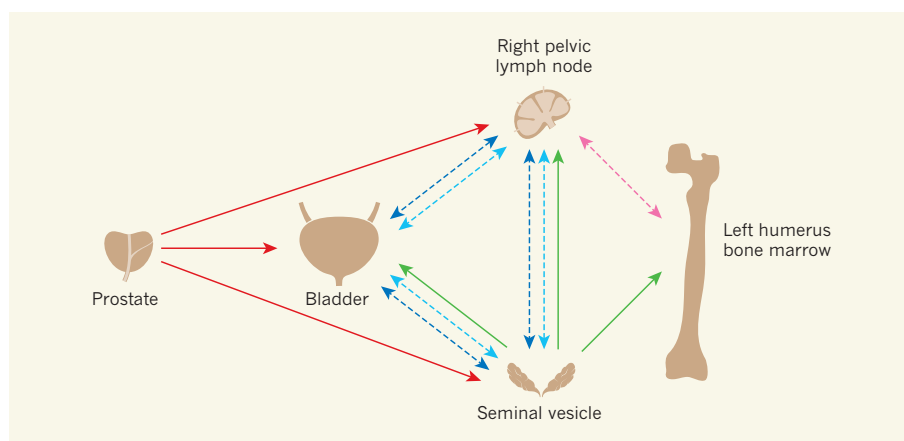
CANCER

The complex seeds of metastasis

Analyses of prostate–cancer metastases reveal a complex cellular architecture, and show that secondary sites can be seeded by multiple cell populations derived from both the primary tumour and other metastases. SEE LETTER P.353

MICHAEL M. SHEN

How does cancer metastasize? The classic 'seed and soil' model, first proposed in 1889, posits that metastases arise from rare tumour cells, or seeds, that survive the steps required to escape the primary tumour and disseminate to secondary tissue sites, proliferating in tissues where there is a compatible microenvironment, the soil[1]. Subsequent experimental studies in prostate[2] and other cancers[3] have suggested that the metastatic seeds correspond to single disseminated cells. But in contrast to this simple viewpoint, two papers (one on page 353 of this issue[4] and one in *Nature Communications*[5]) demonstrate that prostate-cancer metastases often display complex and dynamic patterns of evolution, starting from seeds that are composed of several different cells.

Next-generation DNA sequencing technologies have made it apparent that primary tumours are not clonal (consisting of a single population of genetically identical cells). Instead, they are composed of subclones, subpopulations of genetically identical cells that can be distinguished from other subclones by the mutations they harbour. Such subclones compete for dominance during cancer progression, and drug treatment can lead to formerly minor tumour subclones becoming dominant if they are resistant to treatment[6]. Thus, clonal evolution shapes the properties of tumours and can explain their plasticity in response to therapy. Until now, however, clonal evolution has not been explored in detail in the context of metastasis.

Gundem *et al.*[4] carried out whole-genome sequencing on 51 metastases and primary tumours from 10 patients with lethal prostate cancer, taking advantage of tumour samples that were meticulously collected and stored from a rapid-autopsy programme over 20 years. Hong and colleagues[5] report a similar analysis of 26 samples from 4 patients. The groups analysed the resulting data bioinformatically, clustering mutations on the basis of their clonality or subclonal frequency within each sample, and then reconstructing phylogenetic trees to reflect the lineage relationships of the metastases in each patient. In general, the authors found mutations that

Figure 1 | Complex patterns of metastatic seeding. This simplified schematic exemplifies ways in which cells might seed metastases in patients with prostate cancer. Each subclonal cell population — a group of cells harbouring the same set of genetic mutations — that seeds sites of metastasis is represented by a different coloured arrow, with double-headed, dashed arrows indicating that the direction in which seeding occurs is unknown. Gundem *et al.*[4] and Hong *et al.*[5] found that metastases can be seeded not only by subclones from the primary tumour, but also by those from other metastatic sites. Furthermore, Gundem and colleagues find evidence of polyclonal seeding, in which the same sets of subclones seed multiple sites of metastasis (in this example indicated by dark blue and light blue arrows).

are well-known drivers of prostate cancer in the trunks of these trees, consistent with their occurrence in the primary tumour. However, many of the trees are highly branched as a result of the generation of subclones, and many individual branches are associated with the acquisition of potential driver mutations involved in resistance to therapy.

Remarkably, Gundem and co-workers provide unequivocal evidence that two or more subclones had seeded the same site (a phenomenon known as polyclonal seeding) in at least one metastasis in five of the ten patients they analysed. In addition, multiple subclones were shared between such polyclonal seeds for two or more metastases in each of these five patients, suggesting that these subclones might functionally cooperate with one another to promote metastatic progression. Furthermore, eight of the ten patients displayed metastatic cross-seeding, in which subclones within a metastasis originated from another metastatic site, rather than from the primary tumour (Fig. 1). Such sequential cross-seeding could occur linearly from metastasis to metastasis; in a branched pattern with one metastasis seeding two or more others; or linearly followed

by branching. Hong *et al.* also find evidence of cross-seeding, and describe distinct temporal waves of metastatic seeding from the primary tumour.

The studies also provide insights into the molecular pathways by which prostate tumours acquire resistance to therapy. Because prostate tumours initially depend on androgen hormones such as testosterone, androgen deprivation by chemical or surgical castration is highly effective therapeutically. However, prostate cancer can recur if the tumour becomes 'castration-resistant', which usually occurs through mutations that upregulate androgen-receptor-pathway activity. Gundem and colleagues report that different mutations that each promote castration resistance coexist in distinct subclones within a patient. Furthermore, Hong and colleagues find that blood samples taken at the time of death still contain clones from the primary tumour, several years after its surgical removal. This indicates that circulating tumour cells with the potential to seed metastases persist in the long term.

The heterogeneous (mixed) clonal composition of the metastases described in these two reports raises the issue of whether clonal

heterogeneity might be intrinsic to the metastatic process. Indeed, a mouse model of small-cell lung cancer also shows evidence of polyclonal metastatic seeding[7]. Such seeding might be favoured when two or more distinct tumour subclones cooperate to promote their mutual growth and survival[8–10]. Furthermore, analyses of circulating tumour cells in breast cancer show that metastatic seeding is frequently mediated by small clusters of tumour cells containing multiple clones, rather than by single cells[11].

Taken together, the current studies might explain why, given the prevalence of circulating tumour cells in patients with solid tumours, successful metastasis is relatively rare — metastasis may be facilitated by seeding by cell clusters containing cooperating clones with distinct properties. If so, it is attractive to speculate that disseminated single cells could remain dormant until reawakened by interaction with a cooperative metastatic cell arriving at the same secondary site. Such a model

has the potential to revise our conception of the properties of tumour-initiating cells, as well the metastatic niche, and may have implications for therapeutic strategies. For example, understanding the signalling pathways that mediate such clonal cooperativity may lead to effective therapies using drugs that target these pathways.

Future advances in understanding early events in metastatic seeding will require functional analyses to investigate molecular mechanisms. At present, however, the availability of suitable model systems for such studies is limited. Advances might come from the use of lineage tracing to follow metastasis in genetically engineered mice[12], alongside mathematical methods that assess clonal relationships from genomic data. The development of these and other experimental approaches will undoubtedly accelerate our understanding of the complexity of metastasis. ■

Michael M. Shen *is in the Departments of*

Medicine, of Genetics and Development, of Urology and of Systems Biology, and at the Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, New York 10032, USA.
e-mail: mshen@columbia.edu

1. Fidler, I. J. *Nature Rev. Cancer* **3,** 453–458 (2003).
2. Liu, W. *et al. Nature Med.* **15,** 559–565 (2009).
3. Sethi, N. & Kang, Y. *Nature Rev. Cancer* **11,** 735–748 (2011).
4. Gundem, G. *et al. Nature* **520,** 353–357 (2015).
5. Hong, M. K. H. *et al. Nature Commun.* http://dx.doi.org/10.1038/ncomms7605 (2015).
6. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. *Nature* **501,** 338–345 (2013).
7. McFadden, D. G. *et al. Cell* **156,** 1298–1311 (2014).
8. Calbo, J. *et al. Cancer Cell* **19,** 244–256 (2011).
9. Cleary, A. S., Leonard, T. L., Gestl, S. A. & Gunther, E. J. *Nature* **508,** 113–117 (2014).
10. Marusyk, A. *et al. Nature* **514,** 54–58 (2014).
11. Aceto, N. *et al. Cell* **158,** 1110–1122 (2014).
12. Aytes, A. *et al. Proc. Natl Acad. Sci. USA* **110,** E3506–E3515 (2013).

PLANETARY SCIENCE

A new recipe for Earth formation

Experimental results suggest that if Earth initially grew by the accumulation of highly chemically reduced material, its core could contain enough uranium to drive the planet's magnetic field throughout Earth's history. SEE LETTER P.337

RICHARD W. CARLSON

Determining Earth's bulk composition is difficult because so little of the planet is directly accessible. As a result, most estimates derive from a comparison of Earth rocks with the planet's probable building blocks — the asteroidal bodies sampled by meteorites. New views about the mechanism of planet formation, and their consequences for estimating Earth's composition, are, surprisingly, being driven by observations of planetary systems around stars other than the Sun. On page 337 of this issue, Wohlers and Wood[1] report experiments that explore the consequences of one of these views — that the building blocks of Earth systematically changed composition during its growth. Their results lead to the intriguing conclusion that if Earth formation started with highly chemically reduced building blocks, the planet's metallic core might contain enough uranium to power the convection that creates, and has maintained, Earth's magnetic field for more than 3 billion years.

Most models of planet formation predict that planets grow by the accumulation of smaller bodies known as planetesimals, measuring about 10–100 kilometres in diameter[2]. Planetesimals that form far from the central star are cold enough to include ices, whereas those formed in the hotter region close to the star are largely composed of mixtures of silicate and iron metal. This simple model fits well with the observation of our Solar System's 'snow line', which separates rocky Mars from the giant gaseous outer planets. This implies that planets grow mostly from material that is formed at similar distances from the Sun, and that planets stay in the positions where they are made.

This comforting view of planetary orbital stability was lost with the detection of 'hot Jupiters'[3] — large gaseous planets in orbit close to their stars. These close-in giant planets may have formed far from their suns, as did Jupiter, but then migrated inwards, probably during the period of planet growth. Building on these observations, a theoretical model for planet formation in the Solar System suggests that Jupiter and Saturn may have migrated inwards until gravitational interaction between the two caused them to retreat outwards to their current positions[4].

Because these giant planets are the gravitational heavyweights of the outer Solar System,

their migration during or after the period of planet growth would have scattered smaller planetesimals from their path. This could have resulted in material being cleared from the area where Mars now exists, explaining the planet's small size. If this migration occurred while Earth was forming, the materials from which Earth grew would have changed from those that formed close to the Sun, near Earth's current orbit, to those that formed far from the Sun and were forced into the inner Solar System as they fled in front of inward-migrating Jupiter.

Wohlers and Wood's experiments explore the effects of Earth growing first from the accumulation of highly reduced material, in which most iron is present as metal or sulfide, followed by the accumulation of more-oxidized material, in which a good fraction of the iron is present as iron oxide incorporated into silicate minerals. Within the meteorites delivered to Earth, both highly reduced and highly oxidized varieties exist, so this compositional variation is not just hypothetical. In addition, the discovery of high concentrations of sulfur on the surface of Mercury[5] (Fig. 1) suggests that Mercury is dominated by reduced material, because sulfur is much more volatile when oxidized than when it is reduced. Given Mercury's proximity to the Sun, one might expect it to have formed hot, in which case it should be depleted in sulfur unless it formed preferentially from reduced material.

At present, Earth is quite oxidized. Under these conditions, many elements, including uranium (U), thorium (Th) and the rare-earth elements — those with atomic numbers 57 to 71, such as neodymium (Nd) and samarium (Sm) — are not at all soluble in iron metal or iron sulfide. In this case, Earth's total inventory of U, Th and the rare-earth elements should be

present in the silicate portion of Earth (the crust and mantle), with only insignificant quantities in the planet's iron-metal core. Wohlers and Wood's experiments explored the partitioning of these elements between iron metal, iron sulfide and silicates at different oxidation states. Under very reducing conditions, such as those that might have been involved in Mercury's formation, their results show that U and the lighter of the rare-earth elements, but not Th, become soluble in iron sulfide, which would have joined the iron metal in Earth's core.

The energy source that drives the generation of Earth's magnetic field in the core has long been a topic of discussion, as has the suggestion that one such energy source could be a moderate concentration of uranium or potassium in the core. However, before Wohlers and Wood's experiments, there was only limited (and controversial) experimental evidence that either uranium or potassium can be incorporated in iron metal at the high temperatures and pressures of core formation.

Because U, Th and most of the rare-earth elements condense at very high temperatures from a gas of solar composition, Earth should in principle have accumulated all of these elements in relative abundances similar to those of the Sun — so Earth's bulk abundance ratios of Th to U (Th/U) and Sm to Nd (Sm/Nd) should be the same as the Sun's. But if the core contains a significant amount of U and Nd, as Wohlers and Wood's results suggest, the silicate Earth should be left with higher Th/U and Sm/Nd ratios than those of the Sun. Estimates of the silicate Earth's Th/U ratio are indeed higher than the solar value, but not by enough to make a convincing case for such a selective incorporation of U into the core.

A more stringent test of whether U was incorporated into the core through the mechanism proposed by Wohlers and Wood might be provided by the isotopic composition of Nd in the silicate Earth. This composition is modified by the radioactive decay of two isotopes of Sm — ^{147}Sm (half-life of 106 billion years) and ^{146}Sm (half-life of 103 million years). The average Nd isotopic composition in the silicate Earth was thought to be similar to that of the Sun, with small deviations caused by the chemical separation of a crust with a low Sm/Nd ratio that left the mantle with an elevated Sm/Nd ratio. This assumption began to be questioned when high-precision Nd-isotope measurements in 2005 showed that both crustal and mantle rocks on Earth have a higher ^{142}Nd/^{144}Nd ratio than the Sun[6] (note that ^{142}Nd is the product of the decay of ^{146}Sm).

Given the short half-life of ^{146}Sm, the elevated ^{142}Nd/^{144}Nd ratio of the silicate Earth reflects a higher than solar Sm/Nd ratio that must have been present in the silicate Earth since shortly after its formation. Earth's core formed within tens of millions of years after the formation of the Solar System, so preferential incorporation of Nd into the core, as suggested by Wohlers
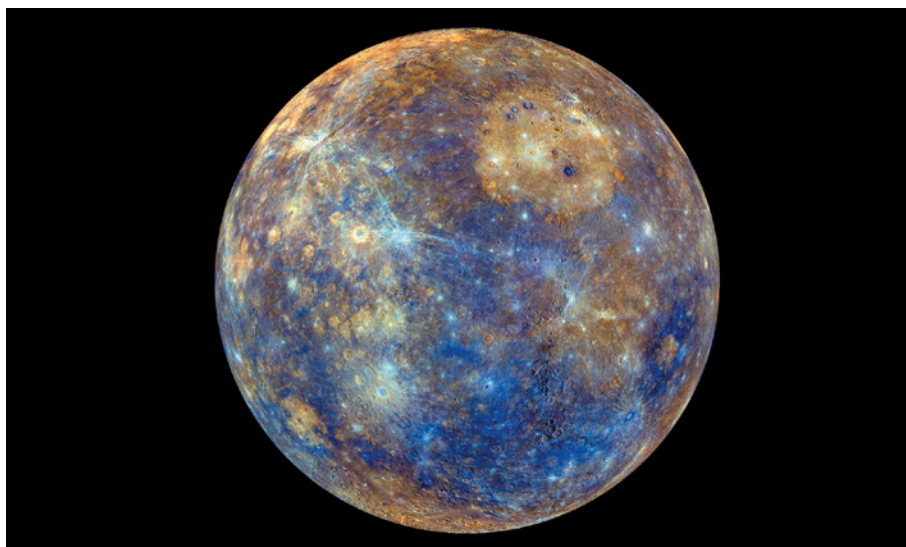
Figure 1 | Mercury's surface. The image is a false-colour mosaic of images of Mercury obtained by NASA's MESSENGER spacecraft. Parts of the planet's surface have atomic ratios of sulfur to silicon that are more than 0.14, compared with values of 0.001 in Earth's mantle[5]. These ratios probably reflect chemically reduced conditions in which sulfur is predominantly contained in iron sulfide, unlike on Earth's surface, where sulfur is mostly in the form of oxidized sulfate. Wohlers and Wood's results[1] indicate that, if Earth's growth initially involved the accumulation of a reduced, Mercury-like body, significant amounts of Earth's uranium may be present in the planet's core.

and Wood's results, would leave the whole silicate Earth with an elevated Sm/Nd ratio that could explain its high ^{142}Nd/^{144}Nd ratio.

To simultaneously satisfy the Th/U and ^{142}Nd/^{144}Nd ratios of the silicate Earth, Wohlers and Wood propose a balance between the amounts of reduced and oxidized materials added to the planet. Although possible, this carefully balanced ratio must also satisfy other potential geochemical consequences of involving highly reduced materials in Earth-formation models — not least, how Earth ended up in its present oxidized state, which it has apparently retained for more than 3 billion years. ∎

Richard W. Carlson *is in the Department of Terrestrial Magnetism, Carnegie Institution for Science, Washington DC 20015, USA.*
e-mail: rcarlson@carnegiescience.edu

1. Wohlers, A. & Wood, B. J. *Nature* **520,** 337–340 (2015).
2. Chambers, J. E. *Earth Planet. Sci. Lett.* **223,** 241–252 (2004).
3. Marcy, G. W. & Butler, R. P. *Astrophys. J. Lett.* **464,** L147 (1996).
4. Walsh, K. J., Morbidelli, A., Raymond, S. N., O'Brien, D. P. & Mandell, A. M. *Nature* **475,** 206–209 (2011).
5. Weider, S. Z. *et al. Earth Planet. Sci. Lett.* **416,** 109–120 (2015).
6. Boyet, M. & Carlson, R. W. *Science* **309,** 576–581 (2005).

CANCER

An extravascular route for tumour cells

Molecular tracing of populations of breast-cancer cells in a primary tumour in mice reveals that two proteins, Serpine2 and Slpi, enable tumour cells to form vascular-like networks, facilitating perfusion and metastasis. SEE LETTER P.358

MARY J. C. HENDRIX

Cancer deaths result primarily from metastases, which comprise mixed populations of tumour cells and which are often resistant to conventional therapies. Metastasis is a complex, multistep process, in which cells escape from the primary tumour

through the vasculature, move from the vascular system into the target organ, and then form secondary tumours[1]. Efforts to study the metastatic properties of different tumour cells that are involved in each aspect of this process have been challenged by inadequate tools — until now. In this issue, Wagenblast *et al.*[2] (page 358) deploy an innovative 'molecular-barcoding'

technology, together with advanced DNA-sequencing techniques, to model the mixed (heterogeneous) populations of cells that comprise breast tumours. This approach enables the authors to study the molecular composition of the different cell populations that contribute to particular aspects of tumour formation and metastasis.

The researchers transplanted aggressive mouse breast-cancer cells into host mice, creating an experimental model of breast-tumour heterogeneity. Each transplanted cell was labelled with a different molecular barcode — a short DNA sequence that has no effect on the cell's function, but which allows the cell and its descendants (known as a clonal population) to be identified through 'next-generation' DNA sequencing. After 24 days, the primary tumours and key organs were collected, and the frequency of each barcoded population within each tissue was quantified by sequencing.

The results revealed that the number of clonal populations within the primary tumour did not correlate with the number of clones in the tumour cells circulating in the blood or within various secondary metastatic growths. Furthermore, different groups of clones contributed to metastases that arose by spread through the lymphatic system compared with spread through the blood. In other words, different clones within the heterogeneous population had distinct properties, such as an ability to dominate the primary tumour, or to contribute to metastatic populations, or to enter the lymphatic or vascular systems.

Wagenblast and colleagues defined the gene-expression profiles associated with different clonal populations, and discovered that those clones that efficiently entered the vasculature expressed two secreted proteins, Serpine2 and Slpi, both of which are anticoagulants[3,4].

Further testing in animal models and *in vitro*, and analyses of other known genes specifically implicated in human lung metastasis, revealed a significant correlation between expression of those genes and Serpine2 and Slpi expression. In addition, overexpression of Serpine2 and Slpi in patients with breast cancer correlated with an increase in metastatic relapse in lungs in which metastasis had been treated previously.

The authors' analyses also revealed that the two anticoagulants probably exert their metastasis-promoting effect by enabling breast-tumour cells to act like cells of the vasculature (Fig. 1). This phenomenon, dubbed vascular mimicry, was first detailed[5] in 1999 in aggressive melanomas, and involves the formation of vascular-like networks by aggressive tumour cells that take on similar characteristics to the endothelial cells that line normal blood vessels. These networks are rich in extracellular matrix and allow perfusion of blood and fluid throughout tumour tissues.

The first reports of vascular mimicry were based on *in vitro* observations in which aggressive-melanoma cells placed in three-dimensional matrices formed perfusable, vascular-like networks similar to the vascular-channel networks lined by tumour cells in patients' tumours. Since this discovery, many studies have contributed insights into the mechanisms that underpin the induction, formation and targeting of vascular mimicry across many types of cancer, including melanomas, sarcomas, carcinomas and glioblastomas[6].

It is notable that stem-like cells from glioblastoma, an aggressive type of brain tumour, can give rise to endothelial-like tumour cells, thus confounding therapeutic strategies targeting the tumour's vasculature[7,8]. Although similar, these tumour-derived cells are not identical to bona fide endothelial cells,

and as such, therapies that target endothelial cells will probably not be effective against them. This has certainly been the case in many trials of inhibitors of blood-vessel formation[9]. However, combinatorial strategies that target both endothelial cells and tumour cells may be a sensible approach.

By identifying Serpine2 and Slpi as drivers of vascular mimicry, Wagenblast and colleagues' study advances our understanding of vascular mimicry in breast-tumour metastasis. The authors hypothesize that the anticoagulant nature of these proteins is key to their metastatic abilities, preventing clotting at the junctions between the newly formed vasculature and the tissue, and they call on a previous study[10] to support their theory. That study, in which human metastatic melanoma was transplanted into mice, made use of Doppler imaging (a technique for tracing blood flow using microbubbles), and showed that there is blood flow between vessels lined by endothelial cells and vascular-mimicry networks lined by tumour cells. The analysis demonstrated that an anticoagulant protein called tissue factor pathway inhibitor 1 is crucial for blood to perfuse into the metastatic tissue in this cancer type. Similar imaging in the current study would have provided stronger evidence for the physiological significance of vascular mimicry in breast-tumour metastasis, and would have allowed the authors to directly test the role of Serpine2 and Slpi from a physiological perspective — perhaps a focus for the future.

As discussed in many reports[11], and further validated by Wagenblast *et al.*, vascular mimicry is associated with a poor clinical outcome. This suggests that vascular mimicry imparts an advantage with respect to tumour perfusion and survival of metastases. Several signalling pathways have been reported to have
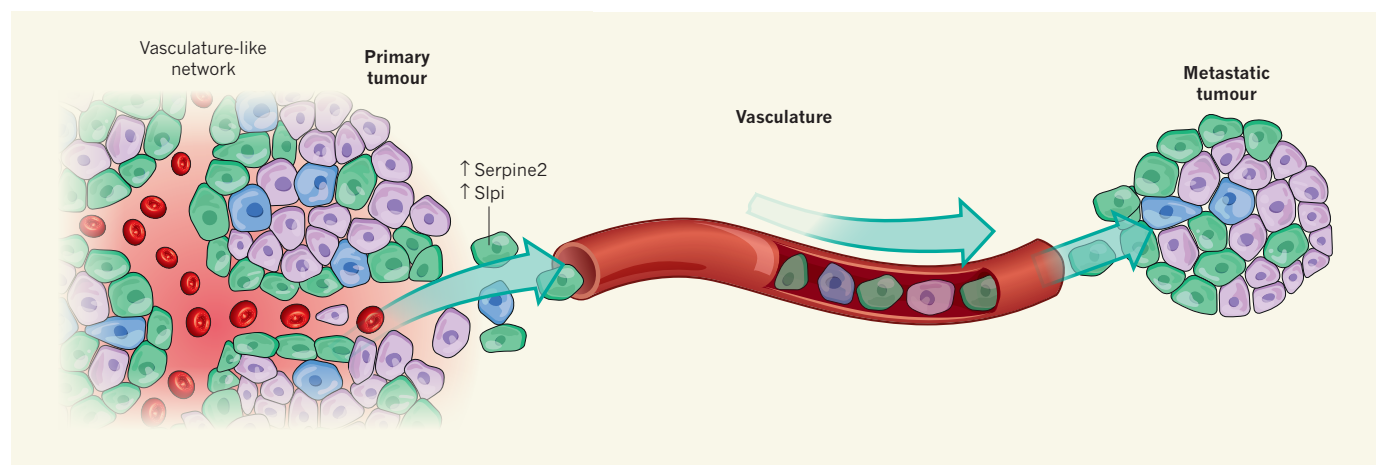
Figure 1 | Vascular mimicry drives metastasis. Tumour cells must undergo several sequential steps to accomplish metastasis: escape from the primary tumour into the vasculature; move through the blood; and escape from the vasculature to become embedded in a distant tissue. Metastasis is promoted by vascular mimicry, a phenomenon in which tumour cells adopt characteristics similar to those of the endothelial cells that line normal blood vessels, and

thereby form vascular-like networks within tumours, and between tumours and blood vessels. Wagenblast *et al.*[2] traced the spread of aggressive mouse breast-cancer cells, and found that two proteins, Serpine2 and Slpi, promoted metastasis by stimulating vascular mimicry — tumour cells expressing these proteins (green) form a vascular-like network that enables other populations of tumour cells (purple, blue) to move easily to secondary sites.

crucial roles in tumour-cell vascular mimicry[9]: these include embryonic and stem-cell pathways; hypoxia-related pathways; and vascular pathways, which we now know to include Serpine2 and Slpi. All these pathways warrant further scrutiny as potential therapeutic targets and diagnostic indicators of metastatic potential.

Wagenblast and co-workers have used an innovative approach to studying the importance of different clones in breast-cancer progression. Further studies should investigate the therapeutic promise of targeting Serpine2 and Slpi. Indeed, the authors' work has

illuminated a crucial initial step in the invasion of tumour cells into the blood that can be used as a model for other cancers, and in the testing of therapeutic strategies. ∎

Mary J. C. Hendrix *is at the Stanley Manne Children's Research Institute, Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, Illinois 60611–2605, USA.*
e-mail: m-hendrix@northwestern.edu

1. Fidler, I. J. *Semin. Cancer Biol.* **21,** 71 (2011).
2. Wagenblast, E. *et al. Nature* **520,** 358–362 (2015).
3. Fayard, B. *et al. Cancer Res.* **69,** 5690–5698 (2009).
4. Sayers, K. T., Brooks, A. D., Sayers, T. J. & Chertov, O. *PLoS ONE* **9,** e104223 (2014).
5. Maniotis, A. J. *et al. Am. J. Pathol.* **155,** 739–752 (1999).
6. Seftor, R. E. B. *et al. Am. J. Pathol.* **181,** 1115–1125 (2012).
7. Ricci-Vitiani, L. *et al. Nature* **468,** 824–828 (2010).
8. Wang, R. *et al. Nature* **468,** 829–833 (2010).
9. Kirschmann, D. A., Seftor, E. A., Hardy, K. M., Seftor, R. E. B. & Hendrix, M. J. C. *Clin. Cancer Res.* **18,** 2726–2732 (2012).
10. Ruf, W. *et al. Cancer Res.* **63,** 5381–5389 (2003).
11. Cao, Z. *et al. Eur. J. Cancer* **49,** 3914–3923 (2013).

ORGANIC CHEMISTRY

Streamlining drug synthesis

Drug manufacture can benefit from flow synthesis, in which raw materials are fed into a sequence of reactors, producing the drug as a continuous output. A flow strategy that capitalizes on solid catalysts has now been realized. SEE LETTER P.329

JOEL M. HAWKINS

The active ingredients of pharmaceuticals are typically structurally complex organic molecules that have precise arrangements of chemical groups. They are made using sequences of organic reactions that build complexity, starting with commercially available compounds and proceeding through typically six to ten chemical steps. Each step is most commonly run in a separate batch reactor — a vessel akin to a laboratory flask, but larger and more sophisticated, to enable the production of active pharmaceutical ingredients (APIs) on kilogram to tonne scales, as required. But on page 329 of this issue, Tsubogo and colleagues[1] describe a different approach to making the anti-inflammatory drug (*R*)-rolipram: a fully continuous

synthesis in which the raw materials flow through a sequence of solid catalysts.

In batch processes for manufacturing chemicals, each step of a synthesis requires specific operations: to combine the reagents under particular reaction conditions, to stop (quench) the reaction, and then to separate the desired product from any components that cannot be taken into the next step, often by isolating and purifying compounds formed at intermediate stages. This can be an arduous process, so there has been increasing interest in flow chemistry and continuous processing[2,3]. In these approaches, starting materials and reagents for a reaction are continuously fed into a reactor such as a tube, so that the resulting intermediate is constantly produced. The products can be directly flowed into the quench, separation and purification steps

to streamline the overall synthesis.

Ideally, different flowing synthetic steps are linked (telescoped), so that intermediates do not have to be isolated, with the ultimate goal being a flow process for the entire synthetic sequence: raw materials and reagents are flowed into a series of reactors, producing the API as the output[4,5]. Such a sequence of flowing steps must be kept in balance to manufacture high-purity APIs; this can be achieved by adding chemical engineering controls and streamlining the underlying chemistry.

As Tsubogo and colleagues describe, flow chemistry can be classified into four types depending on whether the reactants and catalysts are flowed into the reactor together, or whether the flowing reagents pass through solid reagents or catalysts contained in the reactor. This distinction is important when an entire API synthesis is telescoped, because any by-products or excess reagents from one step must be removed before, or tolerated by, the next step. The cleaner the effluent from one step, the simpler the intervening processing required between subsequent steps. And the more tolerant the chemistry of a step is to reagent ratios and to chemical species flowing downstream from previous steps, the simpler are the engineering controls required for that step. This simplicity can be achieved by flowing reagents through solid catalysts (a type IV process as classified by Tsubogo *et al.*), especially if any by-products or unreacted reagents
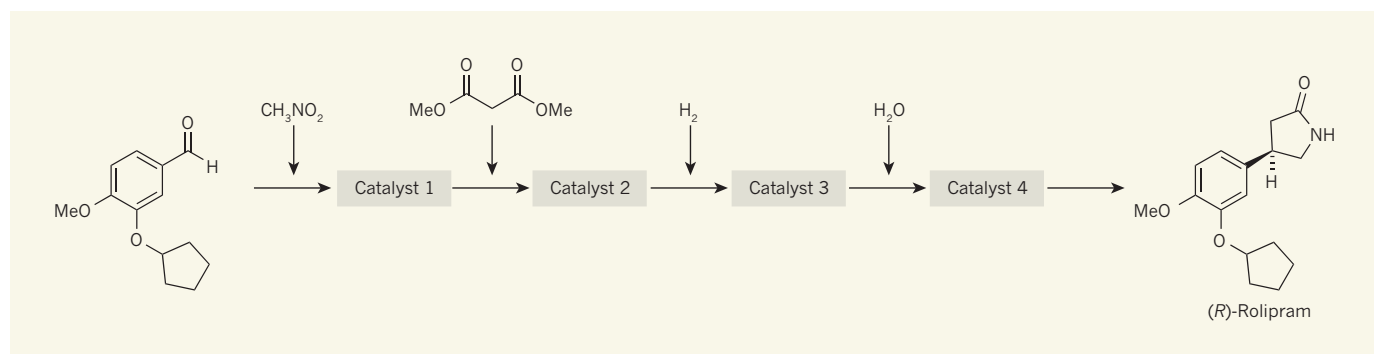
Figure 1 | Flow synthesis of the anti-inflammatory drug (*R*)-rolipram. Tsubogo *et al.*[1] have prepared (*R*)-rolipram by passing a solution of the starting material on the left constantly through a sequence of four solid catalysts, with the added reagents shown above the arrows. The drug emerges as a continuously flowing solution, and as a single product. Me, methyl group.

are innocuous and volatile[6]. Ideally, the solid catalysts stay in the reactor, neither being consumed nor interfering with the chemistry downstream.

Various processes, including oxidations[7], hydrolyses[8] and reactions with hydrogen[9], have previously been reported in which reactants flow through solid catalysts. However, Tsubogo and co-workers are the first to achieve an entire API synthesis by flowing starting materials and reagents through a sequence of such catalysts (Fig. 1). Notably, one of the catalysts was chiral, and so a small amount of this species bound in the reactor tube imparts 'handedness' to a large quantity of intermediate molecules flowing downstream, ultimately contributing to the three-dimensional geometry of the resulting APIs. The researchers demonstrated their system on a laboratory scale, using it to produce gram quantities of drug per day and demonstrating stable operation for at least a week. They are now scaling up the system to the multi-kilogram scale.

The authors' process is inherently modular and flexible, which means that a series of analogues of the original drug can be prepared by simply swapping starting materials or catalysts with structurally different but functionally similar species. Tsubogo et al. demonstrated this by modifying their synthesis to prepare phenibut, a drug from the same family as rolipram.

Kilogram-scale production systems based on Tsubogo and co-workers' system should be small enough to operate in walk-in fume hoods, thus requiring smaller and cheaper infrastructure than is used for conventional batch manufacturing facilities. They might even be small and modular enough to be shipped to a different manufacturing facility, if required for business needs; alternatively, processes could be easily set up to run on an identical but remote sister set-up. Larger quantities of APIs could be achieved by judiciously 'scaling up and scaling out' — increasing reactor sizes within the range that would not require re-engineering and increasing the number of systems run in parallel, within practical constraints. As new APIs become ever more potent, selective and personalized, smaller manufacturing volumes will be needed, making portable, continuous, miniature and modular (PCMM) manufacturing processes such as these particularly attractive (for a discussion of PCMM applied to drug formulation, see ref. 10).

Streamlined drug manufacture using type IV flow systems will require robust catalysts that maintain their chemical activity over time. Alternatively, catalysts that have well-understood deactivation profiles will need to be used at elevated loadings, in tandem with a catalyst-replacement schedule. Most importantly, the range of reactions that are amenable to flowing through solid catalysts must be expanded. This provides a valuable target for future catalyst development, which may in part

be met by solid-supported enzymes (see, for example, ref. 11). ■

Joel M. Hawkins is in Chemical Research and Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, USA.
e-mail: joel.m.hawkins@pfizer.com

1. Tsubogo, T., Oyamada, H. & Kobayashi, S. Nature 520, 329–332 (2015).
2. Pastre, J. C., Browne, D. L. & Ley, S. V. Chem. Soc. Rev. 42, 8849–8869 (2013).
3. Baxendale, I. R. et al. J. Pharm. Sci. 104, 781–791 (2015).
4. Hopkin, M. D., Baxendale, I. R. & Ley, S. V. Chem. Commun. 46, 2450–2452 (2010).
5. Snead, D. R. & Jamison, T. F. Angew. Chem. Int. Edn 54, 983–987 (2015).
6. Battilocchio, C., Hawkins, J. M. & Ley, S. V. Org. Lett. 15, 2278–2281 (2013).
7. Chorghade, R., Battilocchio, C., Hawkins, J. M. & Ley, S. V. Org. Lett. 15, 5698–5701 (2013).
8. Battilocchio, C., Hawkins, J. M. & Ley, S. V. Org. Lett. 16, 1060–1063 (2014).
9. Ouchi, T., Battilocchio, C., Hawkins, J. M. & Ley, S. V. Org. Process Res. Dev. 18, 1560–1566 (2014).
10. Markarian, J. Pharm. Technol. 38(11), 52–54 (2014).
11. Truppo, M. D., Janey, J. M. & Hughes, G. US patent application US20140106413 A1 (2014).

PARTICLE PHYSICS

A weighty mass difference

The neutron–proton mass difference, one of the most consequential parameters of physics, has now been calculated from fundamental theories. This landmark calculation portends revolutionary progress in nuclear physics.

FRANK WILCZEK

Nuclear physics, and many major aspects of the physical world as we know it, hinges on the 0.14% difference in mass between neutrons and protons. Theoretically, that mass difference ought to be a calculable consequence of the quantum theory of the strong nuclear force (quantum chromodynamics; QCD) and the electromagnetic force (quantum electrodynamics; QED). But the required calculations are technically difficult and have long hovered out of reach. In a paper published in Science, Borsanyi et al.[1] report breakthrough progress on this problem.

The difference in mass between neutrons and protons is a very small fraction of their average mass, but the value of that difference is crucial to the structure of the physical world. The neutron, proton and electron masses[2] are 939.56563, 938.27231 and 0.51099906 million electronvolts (MeV), respectively, so the difference between the neutron and proton masses is about 2.53 times the electron mass. Were that mass difference even slightly less than the electron mass, for example if it were one-third of its actual value, then hydrogen atoms would convert into neutrons and neutrinos (through a process called inverse β-decay). Even diminished values for the mass difference that are somewhat larger

The authors' work is a major technical achievement that pushes the envelope of available computer power.

than the electron mass would be catastrophic, because the early Universe would have cooked hydrogen into helium more efficiently than it has, leaving little fuel for hydrogen fusion, the process that sustains normal stars, including our Sun. By contrast, were the mass difference significantly larger than its actual value, then the synthesis of atomic nuclei beyond hydrogen would be difficult or impossible.

Within the currently established framework of fundamental physics, the neutron–proton mass difference is not a primary quantity. It can be calculated in terms of more fundamental inputs. The relevant theories for the calculation are QCD and QED. The formulation of those theories is tight, and their accuracy has been tested rigorously in many applications[3,4]. We can thus identify, with confidence and precision, the fundamental contributions to the neutron–proton mass difference. There are two: electromagnetic interactions and differences in the masses of quarks (the particles that make up hadrons, such as neutrons and protons). Let us discuss them in turn.

If the proton differed from the neutron only in having positive electric charge, and if that charge were roughly uniformly distributed, then the proton would be heavier than the neutron, owing to its additional electrostatic energy. According to Einstein's mass–energy equivalence principle, that extra energy translates into extra mass. More sophisticated estimates, for example using electromagnetism in the context of the quark model of hadrons, lead to the same conclusion.

Fortunately, there is the second contribution. It is convenient, for orientation, to refer

to the quark-model picture of protons and neutrons. According to that picture, the neutron is made from one *u* (up) quark and two *d* (down) quarks, and the proton from two *u* quarks and one *d* quark. According to QCD, the *u* and *d* quarks have exactly the same interaction with gluons, the mediators of the strong force between quarks. But their masses need not be equal, and in fact they are not: the *d* quark is heavier than the *u* quark. Because the neutron differs from the proton in containing a *d* quark in place of a *u* quark, this contribution tends to make the neutron heavier.

Nowadays, we understand that protons and neutrons are complex objects, of which the quark model provides only a crude caricature. For accurate work we must solve the equations of QCD and QED directly, as Borsanyi *et al.* have done, by putting powerful computers to work at clever algorithms. But *u* and *d* quarks still enter as elements of fundamental theory, and they still have different masses and different electric charges, but identical strong interactions. Thus, the two sources of neutron–proton mass difference that figured in our heuristic discussion also underpin the rigorous theory.

If the only place where the mass difference between the *u* and *d* quarks occurred were in the proton–neutron mass difference, we would simply be trading one number for another. But the *u* and *d* (and strange) quarks occur inside dozens of different quark-containing subatomic particles, all of whose masses must be obtained using the same quark-mass values in the QCD and QED equations. Therefore, the quark masses have been determined in a vast number of cases, and the theory is stringent. Borsanyi and colleagues actually calculated the masses of a large number of particles (among them the neutron and proton) and obtain a consistent fit to the observed values.

At this point, an important subtlety deserves mention. Part of the mass of a *u* quark is associated with the energy of its electric field (and likewise for the *d* quark). In fact, that 'self-energy' is, formally, a divergent (non-finite) quantity. Finite, physically meaningful quantities are extracted by calculating finite changes in the self-energy as the quark occurs in different environments, for example when it is bound inside different hadrons. Some physicists find this renormalization theory disconcerting. They would prefer to have a theory in which all fundamental quantities, whether measurable or not, are finite. Disconcerting it may be, but the success of Borsanyi and colleagues' calculations, being firmly based on renormalization theory, profoundly attest to nature's assent to it.

The authors' work is a major technical achievement that pushes the envelope of available computer power. Reading it the other way: with increased computer power, it will be possible to improve on the precision of their result, which is still limited. They have demonstrated convincingly, for the first time, that fundamental theory gives the right sign for the proton–neutron mass difference (not a trivial thing, because, as we have discussed, crude electrostatics gives the opposite result), and have obtained the correct magnitude within a few tens of per cent uncertainty.

From a broader perspective, it is a milestone achievement to include both QCD and QED accurately in the same calculation, because the techniques usually used in those fields — respectively, direct numerical solution (lattice gauge theory) and perturbation theory (Feynman graphs) — are so different. This progress encourages us to predict a future in which nuclear physics reaches the level of precision and versatility that atomic physics has already achieved, with vast implications for astrophysics, and conceivably for technology. We can look forward to much more accurate modelling of supernovae and neutron stars than has so far been possible, and entertain dreams of refined nuclear chemistry, enabling, for example, dense energy storage and ultrahigh-energy lasers. ∎

Frank Wilczek *is at the Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*
e-mail: wilczek@mit.edu

1. Borsanyi, Sz. *et al. Science* **347,** 1452–1455 (2015).
2. Particle Data Group. http://pdg.lbl.gov
3. Kinoshita, T. (ed.) *Quantum Electrodynamics* (World Scientific, 1990).
4. Kronfeld, A. S. & Quigg, C. *Am. J. Phys.* **78,** 1081–1116 (2010).

ECOLOGY

Recovering the potential of coral reefs

An analysis of fish declines in coral reefs shows that simple fishing limits and implementation of marine protected areas can be enough to support recovery of coral ecosystem resilience. See Letter p.341

NICHOLAS K. DULVY & HOLLY K. KINDSVATER

Fishing has transformed today's coral reefs, but its effects can be insidious and hard to detect[1,2]. Defining conservation strategies without an appropriate frame of reference is therefore a serious challenge for coral-reef conservation. In this issue, MacNeil *et al.*[3] (page 341) use a combination of data from protected, near-pristine and fished reefs to document the extent of fish biomass declines that have arisen as a result of fishing. Furthermore, they provide comprehensive evidence that simple forms of fisheries management can successfully restore fish-community biomass.

Conservation biologists typically rely on monitoring over time to track the decline and recovery of biodiversity and ecosystem states[4]. But ecosystem-scale underwater research on coral reefs has been possible only in the past 20 years, so that few suitable time series exist for this purpose. To overcome this challenge, MacNeil and colleagues use a space-for-time substitution[5] to estimate fish biomass from underwater surveys of 832 reefs across the world's tropical oceans. The authors combine data on fish biomass in marine protected areas (MPAs), in which fishing is prohibited, with data from 22 unfished sites that are more than

200 kilometres from the nearest human settlements — the most pristine reefs in the world. This provides an estimate of historical biomass in coral-reef ecosystems on an unprecedented global scale.

The authors find that, on average, there is no more than one tonne of fish biomass in each hectare of protected or near-pristine coral reef, although local ecological conditions can lead to considerable variation (Fig. 1). By comparison, 83% of the fished reefs — both managed and unmanaged — have less than half of this biomass. The range of depletion varies widely, from the most severely degraded reefs in the Caribbean and western Pacific, to almost undetectable depletion in the most remote, least-inhabited islands, such as Pitcairn and Easter islands. The reefs in Guam and Papua New Guinea are near collapse, with only 10% of the historical estimate of fish biomass present.

Although these declines seem dire, an equally important finding is that fisheries management works. This is a message of hope to those working in conservation. Over the past decade, many have given up on fisheries management because it is perceived as being too difficult, expensive or beyond the capacity of academics and non-governmental organizations[6]. Many instead turned to MPAs as a blanket solution to marine-conservation challenges. But to be effective, MPAs need to be

Figure 1 | How overfished is this reef? MacNeil *et al.*[3] show how conservation targets and the recovery rates of key fish groups can be estimated from large-scale comparisons of fish biomass at protected and remote sites.

protected and enforced, which requires them to be large, old and isolated[7]. Effective MPAs can halt declines, but the build-up of biomass to historical levels takes time. MacNeil and colleagues show that recovery takes at least 35 years, twice as long as previous estimates[8]. Patience, persistence and continued financial investment will be essential to the success of the ocean's increasing number of MPAs.

As MacNeil and colleagues recognize, MPAs are simply not an option in areas where people depend on fish from reefs. Coral reefs lie in the waters of more than 100 developing countries, many of which have dense, rapidly growing coastal populations. Enforced MPAs might not be viable because of the burden of displacing fishers, the unknown effects of redistributing fishing and the time it takes for biomass to recover. But the authors show that those reefs

that had some form of management, such as restrictions on fishing equipment, species or access, had 27% more fish biomass than reefs open to fishing.

Even in depleted reef communities, regulations protecting key species can promote ecosystem resilience and recovery. For example, prohibiting specific equipment can allow herbivorous fishes to recover, promoting coral resilience[9]. MacNeil *et al.* take this analysis one step further, comparing MPAs of different ages to predict the recovery speed and sequence of different fish groups following implementation of management measures. Their models predict that species at the base of the food web, including herbivores, will recover rapidly. Some of these low-trophic-level species, such as parrotfishes, recover in a nonlinear manner, reaching the greatest biomass soon after

management is implemented. The researchers predict that these species will be most abundant — and therefore at their most effective for grazing, excavating or scraping away algal overgrowth that limits coral growth — at the time when the reefs recover half of their historic fish biomass.

Piscine predators have historically been the first group to be overfished, and this study shows that they are the last to recover. Because they are almost absent from present-day reefs, their relevance to healthy coral ecosystems is sometimes overlooked. But piscine predators have two essential roles in reef communities. First, they suppress mesopredators such as starfish, preventing trophic cascades that change the dominant reef substrate from hard coral to algal overgrowth. Second, they integrate oceanic and reef food webs, feasting on the planktivorous fishes that vacuum up oceanic zooplankton[10]. Without predatory fishes, reefs are potentially condemned to a state of lowered biomass. Prevention of this negative outcome requires effective fisheries governance, including improved monitoring, equipment restrictions to reduce unintentional catch, and increased transparency in the supply and trade of high-value seafood products[11].

There has been much discussion about coral-reef conservation, but little analysis of the efficacy of alternative management options. Currently, most of the world's coral reefs have little or no management — in part because of the persistent lack of recognition by international development agencies and local governments of the social and economic benefits that small-scale fisheries have for the poorest coastal peoples of the world[11]. MacNeil *et al.* provide definitive confirmation that simple fisheries governance tools, including protected areas and equipment, access and species restrictions, can be effective. If adopted seriously, these measures can secure a sustainable future for coral reefs and the people who depend on them. ∎

Nicholas K. Dulvy and **Holly K. Kindsvater** *are in the Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. e-mail:dulvy@sfu.ca*

1. Wing, S. & Wing, E. *Coral Reefs* **20,** 1–8 (2001).
2. Polunin, N. V. C. & Roberts, C. M. (eds) *Reef Fisheries* (Chapman & Hall, 1996).
3. MacNeil, M. A. *et al. Nature* **520,** 341–344 (2015).
4. Collen, B. *et al. Conserv. Biol.* **23,** 317–327 (2009).
5. Pickett, S. T. A. in *Long-term Studies in Ecology* (ed. Likens, G. E.) 110–135 (Springer, 1989).
6. De Santo, E. M. *J. Environ. Manage.* **124,** 137–146 (2013).
7. Edgar, G. J. *et al. Nature* **506,** 216–220 (2014).
8. Molloy, P. P., McLean, I. B. & Côté, I. M. *J. Appl. Ecol.* **46,** 743–751 (2009).
9. Hughes, T. P. *et al. Curr. Biol.* **17,** 360–365 (2007).
10. Hamner, W. M., Jones, M. S., Carleton, J. H., Hauri, I. R. & Williams, D. M. *Bull. Mar. Sci.* **42,** 459–479 (1988).
11. Sadovy, Y. *Fish Fisheries* **6,** 167–185 (2005).

MAGI3–AKT3 fusion in breast cancer amended

Banerji *et al.*[1] described a novel *MAGI3–AKT3* rearrangement in breast cancer, enriched in triple-negative tumours; the report was highly encouraging as targeted therapies could potentially serve as a new and much needed option to treat this highly aggressive breast cancer subtype. We sought to confirm the presence of this rearrangement in 236 samples of triple-negative breast cancer (TNBC) by using fluorescent *in situ* hybridization (FISH) and reverse transcription–polymerase chain reaction (RT–PCR), and in 84 additional cases from The Cancer Genome Atlas by using FusionSeq. No evidence of the fusion was found in any of the tumours studied. Our study confirms that *MAGI3–AKT3* fusion is not a recurrent event in triple-negative breast cancer, which should be acknowledged before considering the evaluation of targeted therapies in clinical trials. There is a Reply to this Brief Communication Arising by Pugh, T. *et al. Nature* **520,** http://dx.doi.org/10.1038/nature14266 (2015).

TNBC constitutes the majority of breast carcinomas of the basal-like molecular subtype, and is defined by absence of actionable therapeutic targets (ER, PR, HER-2). TNBC patients have a poor response to conventional breast cancer therapies[2] and experience poor survival[3]. As such, molecular elucidation of these tumours is critical in the hopes of developing novel targeted therapies[4].

Discovery of functionally recurrent gene rearrangements is a relatively new approach in breast cancer (for example, MAST kinase and Notch gene families[5]). Banerji *et al.*[1] reported a *MAGI3–AKT3* gene fusion to be present in 7% (5/72) of TNBC. This balanced translocation results in a constitutive activation of AKT kinase, which can be counteracted using small-molecule AKT inhibitors.

We aimed to determine the frequency of *MAGI3–AKT3* fusion in 236 TNBCs represented in high-density tissue microarrays (see Table 1). Following previously described protocols[6–8], FISH was performed using dual colour locus-specific probes for *MAGI3* and *AKT3*. None of the cases showed either *MAGI3* or *AKT3* break-apart or fusion signals. To exclude the possibility of intra-tumour heterogeneity, multiple regions of full tumour sections were screened in a subset of 28 cases, all of which were also negative for break-apart and fusion signals (see Fig. 1).

Figure 1 | Absence of *MAGI3–AKT3* fusion in triple-negative breast cancer. a, Haematoxylin and eosin stained full section of a representative case of triple-negative breast cancer. **b**, **c**, Tissue microarrays (**b**) and multiple regions of the entire section (**c**) were interrogated by FISH (total of 236 cases). **d**, No break-apart signals for *MAGI3* or *AKT 3* were identified. No *MAGI3–AKT3* fusion product was detected by RT–PCR in a subset of 135 cases.

Additionally, shorter primer sequences were designed to test a 187 bp fusion product of intron 9 of *MAGI3* with intron 1 of *AKT3* in archival material. We performed RT–PCR of cDNA in 135 of these cases. No *MAGI3–AKT3* fusion product was detected. Further, we investigated RNA-seq data from 84 TNBC cases from The Cancer Genome Atlas with FusionSeq[9]. We did not find any evidence of *MAGI3–AKT3* gene fusion in these cases either.

Our sample size has sufficient power to detect (with 95% confidence) gene rearrangements that would occur at a frequency of as low as 3%. Based on our results, we can reliably conclude that *MAGI3–AKT3* rearrangement is neither recurrent nor sub-clonal in TNBC. To make the assumption that *MAGI3–AKT3* fusion was a recurrent event in TNBC, Banerji *et al.*[1] interrogated their tumours by using RT–PCR of cDNA followed by Sanger sequencing only. Confirmation at the genomic level by PCR of genomic DNA was performed exclusively in the index case. Hence, we favour the view that, with the exception of the index case, the sequenced RT–PCR products by Banerji *et al.*[1] represent a post-transcriptional fusion event (*trans*-splicing), rather than a true genomic event.

Our patient cohort was mainly Caucasian women, whereas patients of Mexican and Vietnamese decent were studied in Banerji *et al.*[1], raising the possibility that this rearrangement may be population-enriched, a prospect that needs further study.

Methods

Locus specific probes were located at 1p (*MAGI3*: BAC 5′ RP11-1133G15 and 3′ RP11-1008I9) and 1q (*AKT3*: BAC 5′ RP11-931B5 and 3′ RP11-989N14). At least 150 nuclei per case were interrogated in tissue microarrays. In full sections, ~2,000 nuclei per slide were evaluated. FusionSeq[9] is a robust computational tool to detect fusion transcripts in paired-end RNA-seq data[10–12]. Reads were aligned to the human reference genome sequence (GRC37/hg19) using STAR[13]. PCR primers sequences are as follows. *MAGI3* forward: 5′-TGTCCTTGTTCGAGCATCAC-3′, *MAGI3* reverse: 5′-GAGGACACAGTTGCCATTGA-3′, *AKT3* forward: 5′-TGAAAGA AGGTTGGGTTCAGA-3′, *AKT3* reverse: 5′-GCCACTGAAAAGTTGTTGAG G-3′. *PGK* was used as a control gene.

Table 1 | Clinico-pathologic characteristics of 236 triple-negative breast cancers

Patient age	22–92 years	
Tumour size	0.3–7.4 cm	
Number of tumour-type cases†	Invasive ductal carcinoma	218*
	Invasive lobular carcinoma	5
	Metaplastic carcinoma	5
	Other	8
Stage (percentage of cases)	Stage IA	47.1%
	Stage IB	0.6%
	Stage IC	0.6%
	Stage IIA	31.4%
	Stage IIB	6.5%
	Stage IIIA	4.6%
	Stage IIIB	2.0%
	Stage IIIC	4.0%
	Stage IV	2.6%
	Not available	0.6%
Ki-67 (proliferation index)	High (≥ 10%)	219 cases
	Low (< 10%)	17

Weill Cornell cohort (*n* = 153) and University Hospital Zürich cohort (*n* = 83).
†Includes 6 cases from recurrence or metastases, as follows: 1 case of chest wall recurrence; 3 cases of ipsilateral lymph node metastases; 1 case of upper arm metastasis; and 1 case of femoral metastasis (Weill Cornell cohort).
*Includes two patients who had bilateral tumours (Weill Cornell cohort).

BRIEF COMMUNICATIONS ARISING

Juan-Miguel Mosquera[1,2]*, Sonal Varma[1]*, Chantal Pauli[1,2,3], Theresa Y. MacDonald[1], Jossie J. Yashinskie[1], Zsuzsanna Varga[3], Andrea Sboner[1,2,4], Holger Moch[3], Mark A. Rubin[1,2]§ & Sandra J. Shin[1]§

[1]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, New York 10065, USA.
e-mail: sjshin@med.cornell.edu
e-mail: rubinma@med.cornell.edu
[2]Institute for Precision Medicine of Weill Cornell Medical College and New York-Presbyterian Hospital, New York, New York 10021, USA.
[3]Institute for Surgical Pathology, University Hospital Zürich 8091, Switzerland.
[4]Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York 10021, USA.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

1. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 486, 405–409 (2012).
2. Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A. & Caggiano, V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California Cancer Registry. Cancer 109, 1721–1728 (2007).
3. Pal, S. K., Childs, B. H. & Pegram, M. Triple negative breast cancer: unmet medical needs. Breast Cancer Res. Treat. 125, 627–636 (2011).
4. Craig, D. W. et al. Genome and transcriptome sequencing in prospective refractory metastatic triple negative breast cancer uncovers therapeutic vulnerabilities. Mol. Cancer Ther. 12, 104–116 (2013).
5. Robinson, D. R. et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nature Med. 17, 1646–1651 (2011).
6. Berger, M. F. et al. The genomic complexity of primary human prostate cancer. Nature 470, 214–220 (2011).
7. Perner, S. et al. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res. 66, 8337–8341 (2006).
8. Tomlins, S. A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310, 644–648 (2005).
9. Sboner, A. et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 11, R104 (2010).
10. Pflueger, D. et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 21, 56–67 (2011).
11. Mosquera, J. M. et al. Recurrent NCOA2 gene rearrangements in congenital/infantile spindle cell rhabdomyosarcoma. Genes Chromosom. Cancer 52, 538–550 (2013).
12. Mosquera, J. M. et al. Novel miR143-NOTCH fusions in benign and malignant glomus tumors. Genes Chromosom. Cancer 52, 1075–1087 (2013).
13. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

Pugh *et al.* reply

REPLYING TO J.-M. Mosquera et al., Nature 520, http://dx.doi.org/10.1038/nature14265 (2015)

In the accompanying Comment[1], Mosquera and colleagues analysed *MAGI3–AKT3* fusions in 236 formalin-fixed paraffin embedded (FFPE) triple-negative breast cancer (TNBC) specimens using break-apart fluorescence *in situ* hybridization (FISH) and detected no cases with fusions. In contrast, our previous published report found *MAGI3–AKT3* in 8 of 235 breast cancer samples and 5 of 72 TNBC cases by reverse trancriptase polymerase chain reaction (RT–PCR) using gene specific primers[2].

To address this discrepancy, we analysed *MAGI3–AKT3* fusions using a hybrid capture array, 'ExomePlus', that covers known exons, conserved non-coding regions and intronic regions involved in gene fusions including the ~150 kilobase first intron of *AKT3*. FFPE tumour and normal tissue was available from 3 positive TNBC cases from our original screen, including the index case, BR-M-045, and frozen tissue for BR-M-045. We performed ExomePlus hybrid capture and Illumina sequencing, achieving an average median read coverage of 76× (range of medians 29–144) across intron 1 of *AKT3* (chr1:243,859,018–244,006,427), on DNA from these samples.

We found 4 fusion read-pairs within intron 1 of *AKT3* confirming the existence of the *MAGI3–AKT3* fusion in genomic DNA obtained from frozen tissue of the index case, BR-M-045 (Fig. 1)[3]. In contrast, we failed to detect the fusion event in any tumour or normal genomic DNA obtained from FFPE tissue, including the BR-M-045 case. Screening additional DNA from 370 breast tumours, including 280 frozen tumours and 90 FFPE samples, and 372 normals (366 paired samples), also failed to find evidence of the fusion in any of these 370 tumour DNA samples at a threshold of three read pairs.

Comparison of the relative allelic fraction of the fusion event to the median allelic fraction of somatic mutations (Fig. 2)[4], suggests that the *MAGI3–AKT3* fusion event in BR-M-045 may represent a sub-clonal population of tumour cells[5]. Our initial positive observations might therefore be explained by intra-tumour heterogeneity, as well as by rare contamination with the fusion cDNA—that is, we observed the fusion in 4% of cases and in 0 controls, but we analysed only 12 negative controls. In retrospect, any such study, even by a straightforward method such as PCR, would be better powered by using a number of controls equal to the number of experimental samples.

We conclude that, although the *MAGI3–AKT3* fusion occurs in at least one breast cancer case, the overall prevalence is lower than our original estimate. Indeed, the data from Mosquera and colleagues and our validation data, suggest a prevalence of <1%. This rare alteration is oncogenically transforming and its activity is sensitive to Akt inhibition, highlighting the potential of rare genome alterations in breast cancer therapy. Furthermore, the *AKT3* pathway may be important in breast cancer in light of the overexpression of *AKT3* that is observed in basal-like breast cancers[6], the expression subtype corresponding to TNBC, and it remains possible that the *MAGI3–AKT3* translocation will be observed in other cancer types by genomic studies. This Reply has been written on behalf of the original author list[2], T. J. Pugh was not a co-author of the original submitted manuscript but led the analysis of the 'ExomePlus' data.

Trevor J. Pugh[1], Shantanu Banerji[2] & Matthew Meyerson[3,4]
[1]Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada.
[2]Department of Internal Medicine, Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba R3E 0V9, Canada.
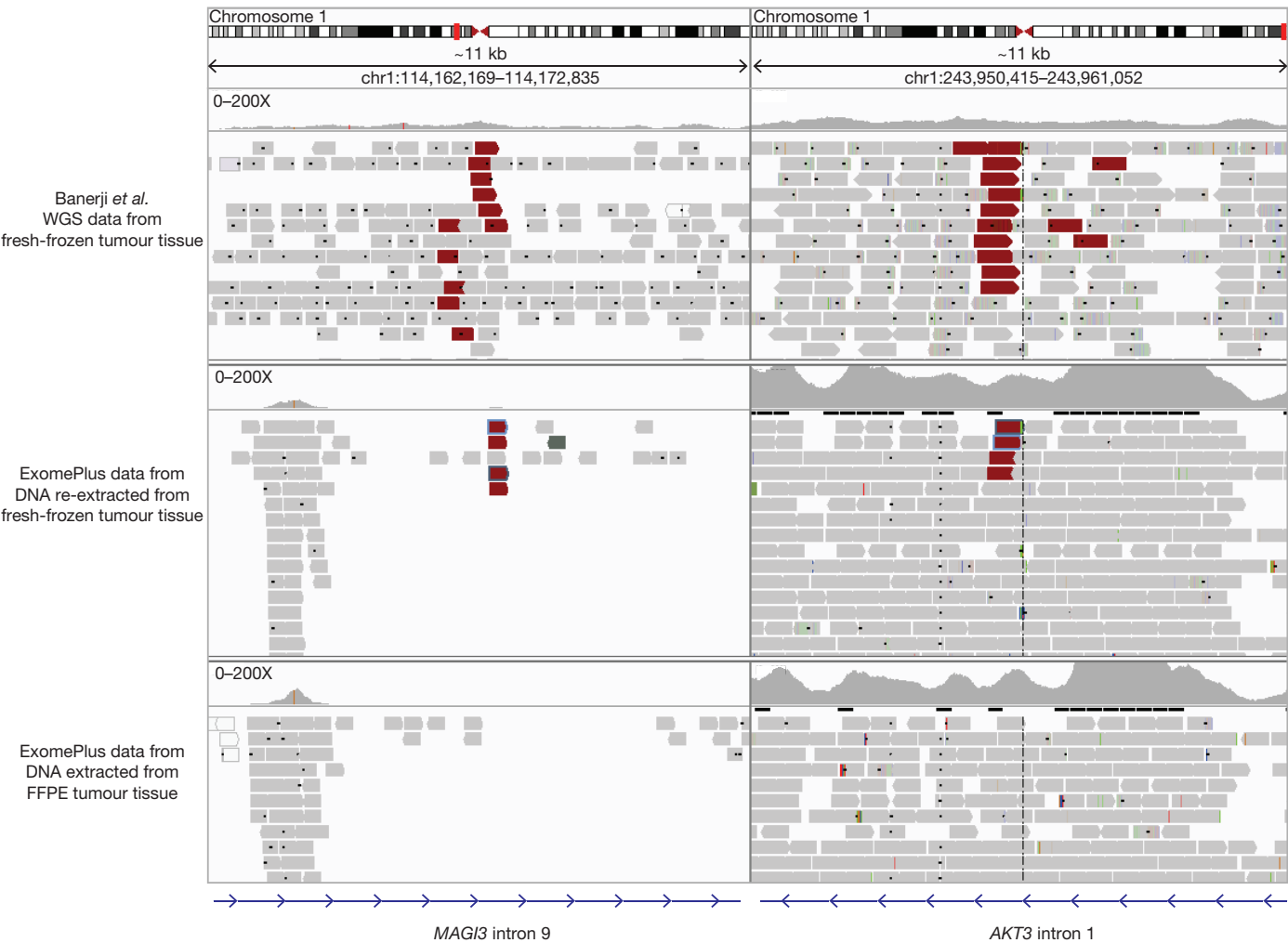
Figure 1 | Integrated Genome Viewer view of *MAGI3–AKT3* translocation sites (*MAGI3* intron 9 and *AKT3* intron 1) from whole genome sequence (WGS) data (top section), ExomePlus capture of frozen tumour (middle section) and ExomePlus capture of FFPE tumour (bottom section). Maroon indicates reads from fusion. Note the absence of coverage of the fusion region for *MAGI3* in the FFPE tumour.
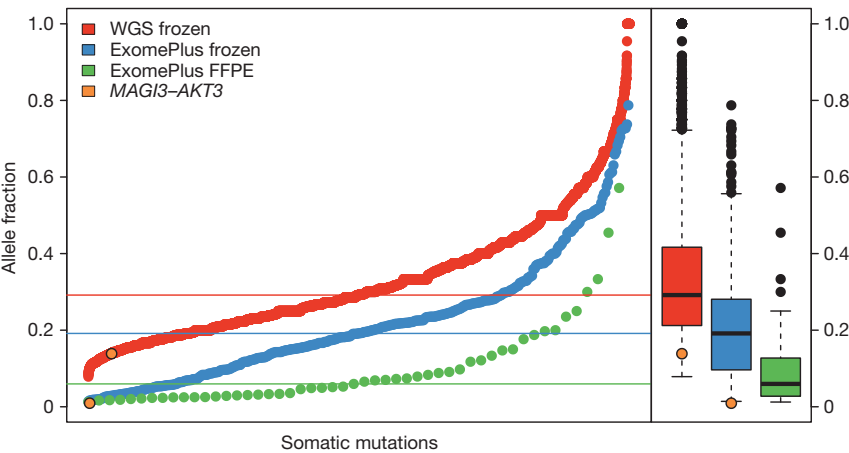
Figure 2 | Allelic fraction detected for somatic mutations in whole genome sequencing of frozen tumour, ExomePlus capture of frozen tumour, and ExomePlus capture of FFPE tumour, compared to allelic fraction of the *MAGI3–AKT3* fusion. Note that capture of the fusion may be less efficient given smaller regions for hybridization to capture probes.

BRIEF COMMUNICATIONS ARISING

3Department of Medical Oncology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.
4Cancer Program, Broad Institute, Cambridge, Massachusetts, 02142, USA.
e-mail: matthew_meyerson@dfci.harvard.edu

1. Mosquera, J.-M. *et al. MAG13–AKT3* fusion in breast cancer amended. *Nature* **520,** http://dx.doi.org/10.1038/nature14265 (2015).
2. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).
3. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnol.* **29,** 24–26 (2011).
4. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31,** 213–219 (2013).
5. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
6. The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

ANALYSIS

A resource for cell line authentication, annotation and quality control

Mamie Yu[1]*, Suresh K. Selvaraj[1]*, May M. Y. Liang-Chu[1], Sahar Aghajani[2], Matthew Busse[2], Jean Yuan[2], Genee Lee[1], Franklin Peale[3], Christiaan Klijn[2], Richard Bourgon[2], Joshua S. Kaminker[2] & Richard M. Neve[1]

Cell line misidentification, contamination and poor annotation affect scientific reproducibility. Here we outline simple measures to detect or avoid cross-contamination, present a framework for cell line annotation linked to short tandem repeat and single nucleotide polymorphism profiles, and provide a catalogue of synonymous cell lines. This resource will enable our community to eradicate the use of misidentified lines and generate credible cell-based data.

The lack of standardization of cell line nomenclature in biological research leads to cell line misidentification, cross-contamination and poor annotation, ultimately affecting scientific reproducibility[1–7]. Cell lines are typically named by the scientist who derived them and only recently have recommendations been proposed[8]. Metadata associated with cell lines also suffers from a lack of consistent and controlled biomedical vocabularies[9,10]. In addition, cell line names are often published with inconsistent syntax and capitalization in the literature as well as in the catalogues of cell line repositories. Figure 1a shows the number of articles in PubMed identified when searching for a selection of cell lines using slight variations of spelling or punctuation in the cell line name. For example, the term 'SK-BR3' identified only 81 related articles, while the term 'SKBR3' identified 645 articles. In this scenario only 5–38% of relevant articles are retrieved, depending on which term is used to search PubMed.

Inconsistent cell line naming also has a significant impact on integrating cell line data for analysis. This has become more apparent in recent years as larger data sets associated with cell line collections become available. For example, comparison of the Sanger[11] ($n = 702$) and the Cancer Cell Line Encyclopedia (CCLE)[12] ($n = 1,046$) cell lines identified 454 common cell lines, of which 59 (13%) of the names are discordant, making cross-referencing these data sets labour intensive and potentially error-prone (Fig. 1b, Supplementary Tables 8 and 9). The most common variations within this analysis are shown in Fig. 1c and often occur in various combinations within the same name (for example, Panc-03-27 and Panc 03.27).

In addition to discrepancies with naming, cell line attributes such as tissue, species, disease type, and pathology are not typically defined using controlled vocabularies. This is apparent even in a resource such as the Cell Line Knowledgebase (CLKB)[9], which draws from ATCC and HyperCLDB[13] to provide a centralized knowledgebase for cell line information. Such variability associated with vocabulary for tissue, cell type and patient diagnosis is commonplace. For example, Supplementary Table 1 lists the different terms which we mapped to 'adenocarcinoma' from source descriptions of tissue diagnosis across multiple databases. All told there are 80 different terms in this field used to describe various samples as adenocarcinoma. To address this problem, we built a framework for describing cell lines available from academic and commercial sources (see Methods). The approach described is largely focused on human oncology cell lines, but can easily be applied to other human and animal cell lines. Within this framework,

each cell line is annotated with uniform baseline categorical data using controlled vocabularies. Supplementary Table 2 lists full annotations for 3,587 cell lines which serves as a foundation for annotation of other cell lines.

Cross-contamination of human cell lines with other human cell lines is a widely acknowledged problem, yet only a minority of scientists confirm the identity of their cell lines or perform adequate quality control for contaminants[14]. Analysis of short tandem repeats (STRs) is the standard test for authenticating cell lines as recommended by the American Type Culture Collection (ATCC) Standards Development Organization Workgroup ASN-0002 (ref. 15), although there are acknowledged drawbacks to using STR profiling[16]. What constitutes "identity" is still open to some debate, as heterogeneity occurs when cells are cultured over extended periods of time, subjected to differing culture conditions or are genetically unstable[3,16]. Loss of heterozygosity, microsatellite instability, aneuploidy in cancer cell lines and cross-contamination make validation problematic. Artefacts due to the procedure (for example, stutter) can affect results and incorrect typing of male cell lines as female is common, owing to deletion of the Y copy of amelogenin or complete loss of the Y chromosome[17]. Comparison of STR gender calls to annotated gender calls for cell lines revealed an unexpected high degree of discordance, with 34% of male lines called as female and 1% of female lines being called male (Table 1). Several STR databases exist (ATCC, DSMZ, JCRB, RIKEN, CLIMA, MD Anderson, Sanger) which allow comparison of cell line STRs to databases of STR profiles. None of these provides a fully curated library of cross-referenced STRs for cell lines, and we found instances of the same cell line mapped to different STR (for example, SNG-II, CCD-14Br in DSMZ) as well as the usual nomenclature inconsistencies. To simplify STR comparisons, we curated a reference file of 2,787 unique STRs from a collection of 8,577 STR profiles (see Methods and Supplementary Table 3). This table removes redundancy, but retains subtle STR variants apparent in cell lines from different sources (for example the TH01 and amelogenin (*AMELX*) loci for SK-N-BE(2) seen in Supplementary Table 4). We also noted that there is no standard mathematical comparison of STR profiles. Methods developed by Tanabe & Masters[18,19] can be implemented in different ways[3], which can cause some confusion over what constitutes a 'match'. Supplementary Table 4 shows results from two online STR-matching tools which return identical matches for STR profiles that clearly vary at several loci. In comparison, we implemented the Tanabe algorithm with rules that return a more accurate

[1]Department of Discovery Oncology, Genentech Inc., South San Francisco, California 94080, USA. [2]Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California 94080, USA. [3]Department of Pathology, Genentech Inc., South San Francisco, California 94080, USA.
*These authors contributed equally to this work.

a

Search term	PubMed hits	Per cent of total
SKBR3	645	38
SK-BR3	81	5
SKBR-3	274	16
SK-BR-3	711	42
SKBR3 OR SK-BR-3 OR SKBR-3 OR SK-BR3	1,702	100
MCF7	21,141	93
MCF-7	19,008	83
MCF7 OR MCF-7	22,633	100
MDAMB231	69	1
MDA-MB231	564	8
MDAMB-231	30	0.4
MDA-MB-231	6,741	92
MDAMB231 OR MDA-MB-231 OR MDAMB-231 OR MDA-MB-231	7,336	100

b

c
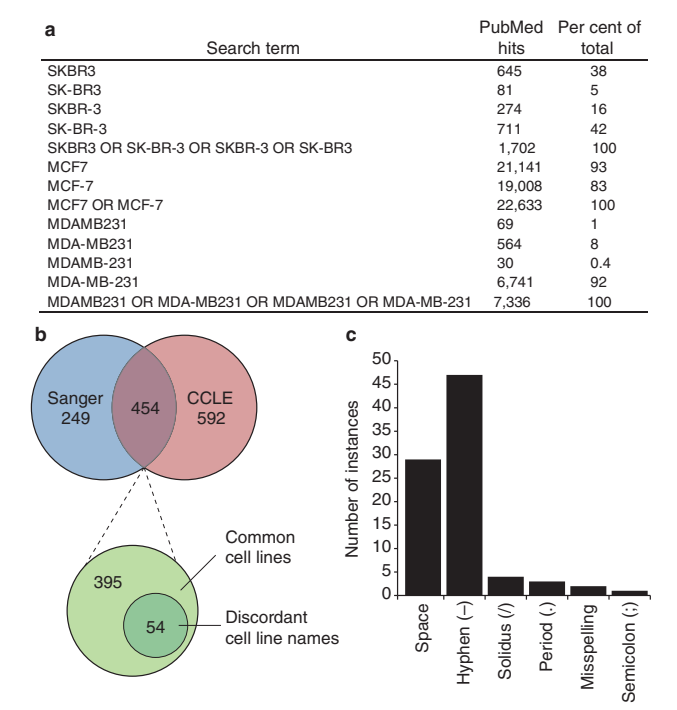
Figure 1 | Inconsistencies in cell line nomenclature. a, PubMed search results using ambiguous cell line terminology. **b**, Venn diagram showing cell lines which are common to the Sanger cell line sequencing project and the Cancer Cell Line Encyclopedia (CCLE). **c**, Graphical representation of the frequency of punctuation/spelling variations which occur in names of cell lines.

Figure 2 | Analysis of STR and SNP fingerprinting of cell lines.
a, Comparison of STR and SNP frequency distributions of pairwise identity alignment scores for 836 lines (see Methods). Heat map colours show joint STR/SNP identity score distribution when computed from true replicate pairs. Reference line shows non-synonymous mean plus 4 standard deviations for STR-based results. **b**, Frequency of synonymous partners detected by STR and SNP analysis. Graph depicting the largest groups of synonymous cell lines (see Supplementary Table 6 for a complete list of synonymous lines). **c**, Graph showing frequency of synonymous partners by tissue/organ of origin.

evaluation of STR matches, which resolves these ambiguities (see Methods).

Single nucleotide polymorphism (SNP) genotyping is another DNA profiling method that can be used to track biosamples[20]. However, an ANSI-approved standard has not been developed for SNP-based cell line authentication. We developed a 48-locus SNP profiling method, using Fluidigm technology, which is a reliable, easy to analyse and cost effective method for quality control of cell line stocks (see Methods). Supplementary Tables 5a and 5b lists the SNP profiles for 1,020 human cancer cell lines using this method whose identity has been verified by STR.

To directly compare the SNP and STR assays, we generated pairwise identity comparisons using 836 cell lines for both STRs and SNPs. This was performed for the standard panel of 8-locus STRs (Extended Data Fig. 1a) and the panel of 16-locus STRs (Fig. 2a) and the 48-locus SNP assay. Biological and technical replicates were highly concordant, supporting the robustness of both assays. Certain derivative cell lines, which represent the same cell line grown in separate culture over extended periods of time, did show greater variation compared with other synonymous partners. For example, HM7 and LS174T were 99% identical by SNP profiling, but only 66% identical by STR. These lines are derivatives and have microsatellite instability, which affects STRs more than SNPs, perhaps explaining the results[21]. Comparison of the HeLa contaminants showed a greater than expected spread of identity scores (Extended Data Fig. 1c), which may be due to the genetically unstable

(aneuploidy, loss of heterozygosity) character of cancer cell lines or poor handling. These data highlight the need for careful and frequent characterization of cell lines, possibly by more than a single method. Our analysis shows a cutoff of 70% identity for 16-locus STRs (85% for 8-loci) and 85% for 48-SNPs is needed to confirm cell line identity. However, due to intrinsic errors of analysing cancer cell lines with either technique[16], we recommend a cutoff of ≥90% identity with either platform to be absolutely certain of a match. Samples below this threshold should be retested, and in cases where a sample fails to match the reference after retesting a new batch should be obtained from the original source.

STR profiling was initially developed as a forensic test for human samples. Forensic STR tests for horses, cattle and canines exist but none

Table 1 | Gender identity for 1,843 cell lines determined by STR compared to annotated gender

STR Call	Annotated	
	Female	Male
Female	855	331
Male	10	600
Total	872	974

that are relevant to cell culture. Primers for mouse STRs are available; however, profiling still remains a challenge, as many mouse cell lines are derived from a handful of inbred strains and thus are indistinguishable, although SNP arrays may be able to resolve this problem[22]. However, the chance of detecting mouse intra-species cross-contamination is low and development of a reliable test is needed. Our hope is that sequencing may become an affordable option to assess human and non-human cross-contamination as costs continue to decline and the genomes of more species are defined.

True synonymous lines are derived from the same patient and have the same DNA profile. Cell lines are also synonymous if they are derived from a parental line *ex vivo* (derivatives) or have been cross-contaminated or misidentified at some point. Identifying synonymous partners (those which share a DNA profile for whatever reason) is critical for basic understanding and interpretation of results. For example, presence of synonymous cell lines could unfairly bias results in studies where panels of cell lines are used to generate correlative data. Despite the excellent efforts of The International Cell Line Authentication Committee (ICLAC)[23], reporting of contaminated and misidentified cell lines is scattered and often inconsistent, thus continued use of cell lines from dubious origins is still evident in the literature. Therefore we sought to create a more comprehensive reference list of synonymous cell lines (see Methods) including legitimate synonymous lines, contaminated and misidentified lines. In total we identified 1,212 cell lines with at least one synonymous partner, including 122 found by STR pairwise comparisons that were not previously reported (Supplementary Tables 6 and 10). We found 27 lines previously reported as cross-contaminated that had unique profiles based on STR analysis and should be regarded as unique (Supplementary Table 7). Cells synonymous with HeLa formed the largest cluster of 143 cell lines whereas 293, HT-29, M14 (MDA-MB-435) and T-24 cell lines represented the largest groups of synonymous partners (Fig. 2b). 22% of synonyms originated from blood-derived cell lines and half of all synonymous partners originated from blood, cervix (HeLa), lung and skin (Fig. 2c). Using this information, we analysed the Sanger and CCLE cell line panels for synonymous partners. In total, CCLE has 69 lines and Sanger contains 6 lines with one or more synonymous partners in their data sets (Supplementary Tables 8 and 9). This table serves as a valuable reference of verified synonymous cell lines as well as a framework for the community to annotate or add further examples as they are identified. We emphasize that many of these synonyms represent legitimate relationships and the provenance of any line should always be researched before use.

Cross-contamination of cell lines occurs through human error such as mislabelling or poor tissue culture technique. Contamination with adventitious organisms (fungi, mould, bacteria) can be readily detected by careful observation or by commercially available tests. It is advisable to test for mycoplasma contamination on a frequent basis as part of good laboratory practice. Here we consider cross-contamination of human cell lines by other established cell lines which often go undetected.

Human (intra-species) cell line contamination is by far the most prevalent and advertised form of contamination as evidenced by the number of cell lines which are HeLa derivatives/contaminants. Cross-contamination seems to occur more frequently in non-adherent (suspension) cell lines (Fig. 2c), but is also prevalent in cultures of adherent cells. The simplest form of misidentification comes from mislabelling, which can be immediately identified if lines are genotyped regularly. Cross-contamination by a small number of contaminating cells is more difficult to detect depending on the ratio of contaminating cells. The reported sensitivity of detection of contaminants is 3% for SNP and 5% for STR, which our own data support. However, the sensitivity of both methods depends on which cells lines are present in the mix, the quality of the data and required detailed review of the raw data (Extended Data Figs 2 and 3).

After the initial event, a low-level contamination can dominate the original culture over time. A contaminant which has a higher rate of
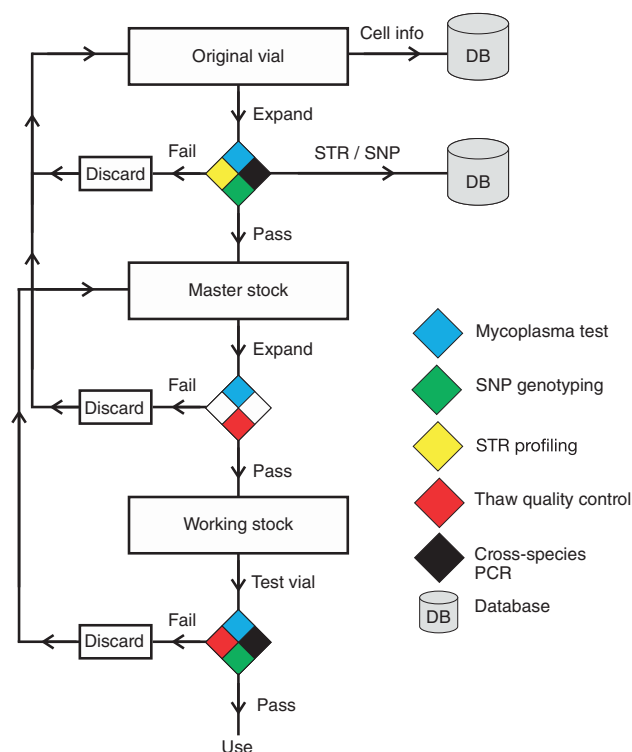
Figure 3 | Flow chart outlining recommendations for maintenance of cell line stocks.

proliferation than its host will overtake the culture. Depending on the rate of growth and the fingerprinting technique, the contamination may not be evident immediately, highlighting the need for continued surveillance. Selective pressures can also select for an underlying contamination. Cells have intrinsic differences in sensitivity to therapeutics as well as antibiotics used for selection of stable transfections. Generating recombinant lines and drug-resistant cells *in vitro* or growth *in vivo* can select for a low-level contaminant present in the original culture. In our experience, the majority of these types of contamination occur in the lab due to inadvertent mix-ups or poor cell culture technique, therefore it is necessary to start with a defined, quality-controlled initial stock of cells and consequently fingerprint the cells once selection is complete.

Non-human (inter-species) cell line contamination has received less attention but is thought to affect approximately 6% of cultures[4]. STR and SNP profiling used to fingerprint human cells do not detect a contaminating sub-population of non-human cells. There are several methods which can be used to detect cross-species contamination but many are not amenable as a standard test in a broad range of laboratories. PCR-based testing has several advantages and can be easily implemented in any laboratory. Dirks and Drexler developed a PCR-based test for rodent mitochondrial DNA[4]; however, we recommend the method developed by Cooper *et al.* and others (see Methods) which detects the cytochrome *c* oxidase subunit I (*COX1*) gene for a broader range of species[24]. Extended Data Fig. 4 illustrates the importance of testing for inter-species contamination. RNA-Seq analysis identified one cell line with an unusually high number of single nucleotide variant calls, which was found to be caused by 21% of the reads mapping to murine sequences. Careful observation of the culture identified two cell morphologies in the cultures, with the smaller, round cells overwhelming the culture after several passages (Extended Data Fig. 4a). The cross-species PCR identified a mix of human and mouse cells which can detect as low as 1% contamination (Extended Data Fig. 4b, c). Contamination was confirmed by detecting human- and mouse-specific CD29 by flow cytometry (Extended Data Fig. 4d).

There is a comprehensive resource of guidelines and good practices for maintenance of quality controlled cell line stocks developed by

experts in cell culture which we cannot cover in detail in this report. Figure 3 outlines a minimal recommended workflow to manage cell line stocks in the average research laboratory (see Methods for details).

In this analysis, we have provided a rich resource of highly curated information for human cell lines with a focus on cancer cell lines. Our analysis of cell line nomenclature attempts to address the issue of ambiguity in biomedical texts. The problem of ambiguity and polysemy of gene names, for example, has been addressed by the HUGO Gene Nomenclature Committee (HGNC) by assigning unique gene symbols, and as journals begin to require correct use of HUGO terms, text mining for gene-related information is gradually improving. In contrast, only recently has a set of guidelines been proposed for cell line terminology[8]. While there have been excellent efforts to define controlled vocabularies and ontologies for existing cell lines[9,13] these have not attempted to reduce the redundancies and complexities perpetuated throughout cell line literature. Our approach was to simplify and unify cell-related information, taking a single name for a cell line and associating it with curated information using a controlled vocabulary. Some discrepancies still exist that need to be resolved by a community-driven consensus to select the most appropriate terms.

Authentication and quality control of cell lines is a unique problem for biomedical science. Almost any other reagent used in science can be defined and characterized with a high degree of certainty so that it can be reproduced with great accuracy. As living, complex biological entities, immortalized cell lines react to their environment and adapt to stresses, leading to appreciable changes over time, probably owing to polyclonality of the original tumour[25,26]. STRs are the current standard for authenticating cell lines and existing databases contain a variety of STRs from different sources. Here we have generated a non-redundant STR database created from publically available STRs and our own data, and have defined simple rules for implementation of an existing matching algorithm that gives an accurate assessment of cell line identity. This provides a foundation for STR comparisons to which more data can be added as more cell lines are profiled.

It is a continuing enigma as to why so many researchers do not authenticate their cell lines. Practices are improving as awareness grows; however, it will require the majority of research institutions, funding agencies and journals to insist upon rigorous cell line authentication before the scientific community views cell line authentication as an essential component of cell-based experimentation[5,6,27,28]. In an attempt to encourage participation in this essential practice, we have presented the methods and data for 48-locus SNP profiling of 1,020 cell lines using Fluidigm technology. Alternatively, the Sanger Institute has made available 97-locus SNP profiles using the Sequenom system for 1,015 cell lines[29]. Although ANSI standards similar to those for STRs have not been developed for SNP profiling yet, our analysis of biologic and technical replicates using both STR and SNP analysis indicates that there is a high degree of confidence that both methods accurately identify cell lines and potential contamination. Together, we hope these alternative and complementary methods for profiling cell lines and biologic samples promote increased surveillance of cell line identity across the community. Balancing the advantages and disadvantages of both methods, we have adopted a policy of deriving STR and SNP profiles for new cell lines. STRs are used to compare with existing external profiles, whereas SNP profiling provides an internal quality control for frequent surveillance of cell lines.

Reporting of synonymous cell lines has increased over the past few years with concerted efforts to identify erroneously labelled cells[23,27]. Here we have collated a resource of more than 1,200 synonymous cell lines. This includes some commercially available derivatives of parental lines, but also identifies unreported synonyms and removes cell lines reported as synonymous that we found to have unique STR profiles. The importance of knowing which lines are identical or mislabelled cannot be underestimated. For example, associative studies across panels of lines should triage cells of common origin to avoid unfair bias. On a more basic level, reporting research using misidentified lines of uncer-

tain origin only serves to confuse the scientific literature. This is probably best illustrated by the MDA-MB-435 cell line used for many years as a model for metastatic breast cancer, but that has the same DNA profile as the M14 melanoma cell line[7]. Evidence that these lines originate from either breast or skin origin has been published, but definitive proof requires access to the original tissue from which the cell line was derived. In the absence of absolute certainty, these lines should not be used in the context of breast or skin cancer research, but perhaps do offer an excellent model for understanding the basic mechanisms of metastasis. Many similar examples are evident in our table where lines with the same profile are stated to be derived from different tissues. Therefore, the combination of the synonym table with defined cell line nomenclature is designed to simplify the process of selecting the appropriate cell lines and avoiding one with uncertain origins.

In conclusion, we have outlined a comprehensive framework for cell line authentication, quality control, annotation and data integration that can be easily adopted, expanded and improved by our community. We have attempted to provide simple solutions to pervasive problems associated with the cultivation of cell lines and sharing of cell-based data, and encourage others to contribute ideas to finally resolve these issues and improve reliability of cell-based research.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell line misidentification: the beginning of the end. *Nature Rev. Cancer* **10,** 441–448 (2010).
2. Editorial. Identity crisis. *Nature* **457,** 935–936 (2009).
3. Capes-Davis, A. *et al.* Match criteria for human cell line authentication: where do we draw the line? *Int. J. Cancer* **132,** 2510–2519 (2013).
4. Dirks, W. G. & Drexler, H. G. STR DNA typing of human cell lines: detection of intra- and interspecies cross-contamination. *Methods Mol. Biol.* **946,** 27–38 (2013).
5. Editorial. Announcement: Reducing our irreproducibility. *Nature* **496,** 398 (2013).
6. Lorsch, J. R., Collins, F. S. & Lippincott-Schwartz, J. Fixing problems with cell lines. *Science* **346,** 1452–1453 (2014).
7. Lacroix, M. Persistent use of "false" cell lines. *Int. J. Cancer* **122,** 1–4 (2008).
8. ICLAC. Naming a Cell Line http://iclac.org/resources/cell-line-names/ (2014).
9. Sarntivijai, S., Ade, A. S., Athey, B. D. & States, D. J. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics* **24,** 2760–2766 (2008).
10. Hunter, L. & Cohen, K. B. Biomedical language processing: what's beyond PubMed? *Mol. Cell* **21,** 589–594 (2006).
11. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
12. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
13. Romano, P. *et al.* Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res.* **37,** D925–D932 (2009).
14. Buehring, G. C., Eby, E. A. & Eby, M. J. Cell line cross-contamination: how aware are mammalian cell culturists of the problem and how to monitor it? *In Vitro Cell. Dev. Biol. Anim.* **40,** 211–215 (2004).
15. Barallon, R. *et al.* Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell. Dev. Biol. Anim.* **46,** 727–732 (2010).
16. Parson, W. *et al.* Cancer cell line identification by short tandem repeat profiling: power and limitations. *FASEB J.* **19,** 434–436 (2005).
17. Santos, F. R., Pandya, A. & Tyler-Smith, C. Reliability of DNA-based sex tests. *Nature Genet.* **18,** 103 (1998).
18. Tanabe, H. *et al.* Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24. *Tiss. Cult. Res. Commun.* **18,** 329–338 (1999).
19. Masters, J. R. *et al.* Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc. Natl Acad. Sci. USA* **98,** 8012–8017 (2001).
20. Castro, F. *et al.* High-throughput SNP-based authentication of human cell lines. *Int. J. Cancer* **132,** 308–314 (2013).
21. Much, M., Buza, N. & Hui, P. Tissue identity testing of cancer by short tandem repeat polymorphism: pitfalls of interpretation in the presence of microsatellite instability. *Hum. Pathol.* **45,** 549–555 (2014).
22. Didion, J. P. *et al.* SNP array profiling of mouse cell lines identifies their strains of origin and reveals cross-contamination and widespread aneuploidy. *BMC Genomics* **15,** 847 (2014).
23. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127,** 1–8 (2010).
24. Cooper, J. K. *et al.* Species identification in cell culture: a two-pronged molecular approach. *In Vitro Cell. Dev. Biol. Anim.* **43,** 344–351 (2007).

25. Masters, J. R. & Stacey, G. N. Changing medium and passaging cell lines. *Nature Protocols* **2,** 2276–2284 (2007).
26. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science **346,** 256–259 (2014).
27. Masters, J. R. Cell-line authentication: end the scandal of false cell lines. *Nature* **492,** 186 (2012).
28. Nardone, R. M. Eradication of cross-contaminated cell lines: a call for action. *Cell Biol. Toxicol.* **23,** 367–372 (2007).
29. Wellcome Trust Sanger Institute. The Cell Lines Project http:// cancer.sanger.ac.uk/cancergenome/projects/cell_lines/about (2015).

Supplementary Information is available in the online version of the paper.

METHODS

Definitions. Synonymous: lines which, by DNA profiling (STR, SNP) have common origins. Lines can be synonymous because they are (1) serial biopsies from the same patient, (2) derivatives from a parental line (drug or clonal selection, transfection etc), (3) misidentified.

Misidentified: a cell line which has a DNA profile that no longer matches the original donor. This can occur by mislabelling or cross-contamination.

No statistical methods were used to predetermine sample size.

Cell line nomenclature, annotation. Cell line information was drawn from cell line repositories (ATCC, DSMZ, JCRB, ECACC) and other sources such as the NCI cell lines and academic institutions. Our initial list contained 6,857 cell lines including duplicates. These were consolidated into a single entry resulting in a final list of 3,587 cell lines. In addition to redundant names, cell lines derivatives were removed (the derivatives wrap up to the cNAME, or the parental cell line). Manual curation of the cell line name and associated information harmonized attributes such as punctuation and capitalization differences between data sources. Inconsistent and often incorrect usage of pathology terms were corrected to terms which adhere to The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)[30]. In situations where cell line names varied between data sources, we attempted to find the original publication to adhere to the author's intent, in cases where this was not possible we used the most common name usage. In cases where nomenclature varied in original publications, a single format was selected and applied to all similarly named lines.

For a cell line to be entered into our database four attributes are required; (1) cell line Name is a unique name identifying the cell line; (2) species is the taxonomic categorization of the organism from which the cell line was derived; (3) primary tissue is the tissue from which the cell originated. This may not be the same as the site of extraction in the case of metastatic samples such as CAL-148 which is a breast cancer cell line extracted from the pleural cavity; (4) tissue diagnosis is the pathology of the sample. Other attributes include: site of extraction, age, gender and ethnicity. In instances where attributes are not known or ambiguous 'unknown' is used until the information is made available.

Each cell line is annotated with the following terms; patient identifier, common cell line name (cName, described below), species, primary tissue and tissue diagnosis. The patient identifier is a unique string that connects cell lines derived from the same patient. The cName is a controlled name for a particular cell line that in most cases matches the spelling and syntax of the first published instance of the particular cell line. The primary tissue and tissue diagnosis terms describe the tissue from which the sample was derived, and the diagnosis of the tissue, respectively. Additional descriptive content can be used to annotate cell lines using controlled vocabularies for fields such as sex, ethnicity or age.

While the primary tissue and tissue diagnosis terms for some cell lines are well documented, there are others for which less is known. This produces a variable level of annotation across cell lines, complicating some analyses. As such, two very simple ontologies were added to the framework to allow straightforward aggregation of samples of interest. The diagnosis ontology is simply a mapping of each tissue diagnosis term to either 'cancer' or 'normal' to more easily compare cancer to normal samples. The tissue type ontology maps each tissue to a more general term, and an example of such a mapping is 'caecum' to 'colon'. While these very simple ontologies have general utility for addressing cancer-focused questions, additional ontologies could very easily be generated to address questions more relevant to other disease areas.

The controlled cell line annotations and the two ontologies have a profoundly useful impact on cell-line based analyses. The controlled vocabularies for all fields are included in their entirety in Supplementary Tables 11–14. This collection of authenticated cell line data will be made available through NCBI's BioProject and BioSample databases, accessible through accession number PRJNA271020, for continued community development and refinement.

cName concept. In the simplest case, cName = cell line name. If derivatives of the parental line are made, these share the cName but have a different cell line name. When two or more cell lines are derived from the same patient, these share the cName if the tissue and diagnosis are identical. If cells are derived from different organs or diseased tissues a separate cName can be issued. In historical cases where two lines are derived from the same patient or cell lines are found to be identical with no history (a possible contaminant), a single cName was chosen when information was available describing the methodology. In cases where cell line origin it is less clear (for example, SK-BR-3/AU565) the lines retain a separate cName and are marked as synonymous in the synonym table.

STR reference database. 8,577 STR profiles were obtained from public databases and generated from our own cell line collection, and pairwise similarity scores (using the Tanabe algorithm[18]) were generated to identify redundancy and synonymous lines. STRs which matched with a score ≥0.9 (90% identical) were first filtered for redundant samples (that is, STR profiles of the same cell line from different sources). Those with identical STR profiles but different cell line names were grouped and used to populate the synonym table, leaving a single STR profile to represent each synonym group in the reference table. This simplifies the output when comparing sample to reference). In cases where STR profiles for synonymous cell lines, derivatives or misidentified lines from different sources were not an exact match (between 90 and 100% match), a single example of each of these were retained in the database to capture this diversity (for example, see the CCRF-CEM cluster in Supplementary Table 3). The final STR reference table contains 2,786 unique profiles.

Synonymous cell line table. Synonymous cell lines were primarily identified by pairwise-analysis of the STRs gathered from multiple sources (Supplementary Table 10). This was cross-referenced with published cell line tables: (1) the Sanger Cell Line resource (http://www.sanger.ac.uk/genetics/CGP/Genotyping/synlinestable.shtml), (2) the ICLAC list of contaminated lines, version 7.2, released 10 October 2014 (refs 23, 31) and Wikipedia (http://en.wikipedia.org/wiki/List_of_contaminated_cell_lines), (3) reported in cell line repositories or the literature. Derivatives of cell lines which are commercially available were retained in this list. Cell lines reported as synonymous which were found to have a unique STR profile compared to the reported contaminant, were excluded from the list. To avoid ambiguity, an STR identity cut-off of 90% was used to call two lines synonymous.

Reporting misidentified cell lines. Misidentified cell lines occur because there was (1) an error at source- the cell line was a contaminant from the outset and the original line never existed or was lost, (2) the original stock exists and is unique, but a contamination subsequently arose and was distributed, and (3) a 'virtual' error occurs when the cell line exists and is unique, but a sample or data handling occurred. With the correct follow-up (that is, repeating/confirming the result by obtaining and testing a fresh sample from the original source) the error type can be determined, and should not be publicized as misidentified unless it is proven to originate at the source.

Cell line STR and SNP profiling. *Short tandem repeat (STR) profiling.* DNA was extracted from cells (Qiagen DNeasy Blood & Tissue (catalogue number 69506)), the concentration determined and normalized to 50 ng ml^{-1}. An aliquot of each was retained for SNP genotyping to identify any sample handling errors. STR analysis was performed by a third party (Genetica DNA Laboratories Inc.) using the PowerPlex 16 HS (Promega Corporation) kit which analyses 16 independent genetic sites specific for human DNA that include the 13 CODIS loci, plus PENTA E, PENTA D and amelogenin. The resulting STR DNA profile report (including allele designations and the raw data of the alleles with their graphic profiles depicting allele peak heights and areas) was used to compare against a curated list of STR profiles.

STR authentication and comparison to reference STRs. For either SNP or STR data, we applied the Tanabe algorithm (or Sørensen similarity index)[18] and computed an identity score for any pair of samples as follows: for each locus at which sample 1 and sample 2 both have called alleles (that is, where neither is a 'no call'), we computed (1) the total number of distinct alleles seen in sample 1, (2) the total number of distinct alleles seen in sample 2, and (3) the number of distinct alleles shared by both samples. Each of the three counts was then summed across all loci, and the identity score was defined as $2 \times$ shared/(total 1 + total 2). The identity score is 0 if and only if no common alleles are seen at any locus; it is 1 if and only if the exact same alleles are seen in both samples at all loci. Note that this approach does not assume diploid genomes or biallelic markers, nor does it require that the same set of markers be available for every pair of samples.

After comparing the query profile against all STR profiles, the match is used to categorize the reference profiles as close matches (>90%) and poor matches (80–90%) to the query STR profile.

Comparison of STR and SNP profiles. Pairwise alignment scores were calculated for 836 cell lines (Fig. 2a). Heat map colours show joint STR/SNP identity score distribution when computed from true replicate pairs (48 replicate pairs for the STR assay and 2,862 replicate pairs for the SNP assay). Identity scores are computed using the Tanabe algorithm for both 16-locus STR and 48-locus SNP genotype results. Total number of comparisons was 349,030 (348,953 non-synonymous and 77 synonymous pairs of cell lines). Univariate distributions for 16-locus STR and 48-locus SNP identity scores and a comparison of 8-locus STR and 48-locus SNP genotype are shown in Extended Data Fig. 1. For plotting purposes, a random subset of 25,000 non-synonymous pairs is displayed. Synonymous cell line pairs are well separated from the large cluster of non-synonymous pairs, but only a subset of synonymous pairs achieve identity scores similar to those typically seen for true replicate pairs.

SNP fingerprinting. SNP genotypes are performed each time new stocks are expanded for cryopreservation. Cell line identity is verified by high-throughput

SNP genotyping using Fluidigm multiplexed assays[32]. SNPs were selected based on minor allele frequency and presence on commercial genotyping platforms. SNP genotyping reactions were setup according to manufacturer's instructions using the single target amplification method. Genotyping was performed on the Fluidigm 48.48 Dynamic Arrays and fluorescence intensity was measured on the Biomark HD System. Data analysis was done with Fluidigm SNP Genotyping Analysis v4.0.1 with a confidence threshold of 95. All genotyping calls were manually checked for accuracy and ambiguous data points were scored as no calls.

SNP profiles are compared to SNP calls from available internal and external data (when available) to determine or confirm ancestry. In cases where data are unavailable or cell line ancestry is questionable, DNA or cell lines are re-purchased to perform profiling to confirm cell line ancestry. SNPs analysed: rs11746396, rs16928965, rs2172614, rs10050093, rs10828176, rs16888998, rs16999576, rs1912640, rs2355988, rs3125842, rs10018359, rs10410468, rs10834627, rs11083145, rs11100847, rs11638893, rs12537, rs1956898, rs2069492, rs10740186, rs12486048, rs13032222, rs1635191, rs17174920, rs2590442, rs2714679, rs2928432, rs2999156, rs10461909, rs11180435, rs1784232, rs3783412, rs10885378, rs1726254, rs2391691, rs3739422, rs10108245, rs1425916, rs1325922, rs1709795, rs1934395, rs2280916, rs2563263, rs10755578, rs1529192, rs2927899, rs2848745, rs10977980.

Fluorescence activated cell sorting (FACS) analysis of CD29. Cells were dissociated using Cell Dissociation Buffer, Enzyme-Free Hank's (Life Technologies, 13150-016). Approximately 1×10^6 cells were collected, washed twice with ice cold staining buffer (PBS, 5% FBS). Cells were co-stained on ice for 20 min with conjugated antibodies: CD29 mouse anti-human monoclonal antibody, Alexa Fluor 488 (Life Technologies, CD2920), at 1:100 dilution and CD29 hamster anti-mouse/rat monoclonal antibody, allophycocyanin (Life Technologies, A14888), at 1:200 dilution, in 100 μl staining buffer at 4 °C. The cells were washed twice with ice cold staining buffer, re-suspended in 300 μl of staining buffer + 0.1 mM Hoechst and incubated at 4 °C for 15 min before sorting. Cells were sorted using the BD LSRII flow cytometer collecting 200,000 gated events.

Cytochrome _c_ oxidase I gene (COI) multiplexed PCR. This method was developed by Cooper _et al._ and others[24,33–35]. Species-specific primer sequences were designed by Parodi _et al._[33] and Cooper _et al._[24]. Multiplexed primer concentrations were based on Cooper _et al._, mixed with 25 ng DNA and JumpStart REDTaq Ready Mix (Sigma-Aldrich) to a final volume of 50 μl. Multiplex cycling conditions: One cycle of 95 °C for 3 min; 30 cycles of 95 °C for 30 s, 60 °C for 15 s,

72 °C for 30 s; 1 cycle of 72 °C for 7 min; and indefinite hold at 4 °C. PCR products were visualized on 4% precast gels stained with ethidium bromide (Invitrogen).

Guidelines for maintaining the integrity of cell line stocks. Quality controls are required at each step to avoid human error and contamination. Upon receipt of a cell line (Original Vial) it is expanded, preferably in a separate quarantine facility dedicated to accessioning new cell lines. Information for the cell line is stored in a database using the defined nomenclature and ontology outlined previously. Cells are tested for mycoplasma and cross-species contamination, and baseline STR and SNP fingerprint profiles are generated to confirm identity. The expanded cells are stored as master stocks to maintain a low-passage source of the cell line. These are then expanded, tested for mycoplasma, and banked as working stocks. A test vial of the working stock is thawed and expanded to confirm cell viability (thaw quality control), and mycoplasma, cross-species contamination and SNP genotyping quality controls are performed before these stocks are used/distributed. New working stocks are generated from the existing working stock for up to 20 passages past the master stock, after which a master stock vial is expanded to generate a new working stock. Failure of quality controls at any stage requires re-testing as false-positives or sample mix-ups can occur. If confirmed, a new vial of the previous stock should be obtained and re-tested. After the initial expansion, all subsequent re-expansions of cell line stocks, or routine quality control of cell lines, are monitored using the SNP platform. Linking cell line annotations (using defined terms) with STR/SNP profiles in a database provides the foundation to associate any cell-based data with the cell line of origin thus facilitating data integration and comparison.

30. Centers for Disease Control and Prevention. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). (2011).
31. ICLAC. Database of Cross-contaminated or Misidentified Cell Lines http://iclac.org/databases/cross-contaminations/ (version 7, 2, released 10 October 2014).
32. Wang, J. _et al._ High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. _BMC Genomics_ **10,** 561 (2009).
33. Parodi, B. _et al._ Species identification and confirmation of human and animal cell lines: a PCR-based method. _Biotechniques_ **32,** 432–434,–436, 438–440 (2002).
34. Steube, K. G., Meyer, C., Uphoff, C. C. & Drexler, H. G. A simple method using beta-globin polymerase chain reaction for the species identification of animal cell lines–a progress report. _In Vitro Cell. Dev. Biol. Anim._ **39,** 468–475 (2003).
35. Hebert, P. D., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. _Proc. R. Soc. Lond. B_ **270,** 313–321 (2003).

ARTICLE

Crystal structures of the human adiponectin receptors

Hiroaki Tanabe[1,2,3,4], Yoshifumi Fujii[1,4], Miki Okada-Iwabu[5,6]*, Masato Iwabu[5,6,7]*, Yoshihiro Nakamura[1,3,4]*, Toshiaki Hosaka[1,3], Kanna Motoyama[1], Mariko Ikeda[1,3], Motoaki Wakiyama[1,3], Takaho Terada[1,4], Noboru Ohsawa[1,3], Masakatsu Hato[1,3], Satoshi Ogasawara[8], Tomoya Hino[8,9], Takeshi Murata[1,8,9,10], So Iwata[1,8,9,11,12,13], Kunio Hirata[13], Yoshiaki Kawano[13], Masaki Yamamoto[13], Tomomi Kimura-Someya[1,3], Mikako Shirouzu[1,3], Toshimasa Yamauchi[5,6,14], Takashi Kadowaki[5,6] & Shigeyuki Yokoyama[1,2,4]

Adiponectin stimulation of its receptors, AdipoR1 and AdipoR2, increases the activities of 5′ AMP-activated protein kinase (AMPK) and peroxisome proliferator-activated receptor (PPAR), respectively, thereby contributing to healthy longevity as key anti-diabetic molecules. AdipoR1 and AdipoR2 were predicted to contain seven transmembrane helices with the opposite topology to G-protein-coupled receptors. Here we report the crystal structures of human AdipoR1 and AdipoR2 at 2.9 and 2.4 Å resolution, respectively, which represent a novel class of receptor structure. The seven-transmembrane helices, conformationally distinct from those of G-protein-coupled receptors, enclose a large cavity where three conserved histidine residues coordinate a zinc ion. The zinc-binding structure may have a role in the adiponectin-stimulated AMPK phosphorylation and *UCP2* upregulation. Adiponectin may broadly interact with the extracellular face, rather than the carboxy-terminal tail, of the receptors. The present information will facilitate the understanding of novel structure-function relationships and the development and optimization of AdipoR agonists for the treatment of obesity-related diseases, such as type 2 diabetes.

Adiponectin (encoded by *ADIPOQ* in humans)[1–4] is an anti-diabetic adipokine. Plasma adiponectin levels are reduced in obesity and type 2 diabetes[5], while the replenishment of adiponectin reportedly ameliorated glucose intolerance and dyslipidaemia in mice[6–8]. These beneficial effects of adiponectin are likely to be exerted, at least in part, by the activation of AMPK[9–11] and PPAR-α[12,13].

We previously reported the expression cloning of the complementary DNAs encoding adiponectin receptors 1 and 2 (*ADIPOR1* and *ADIPOR2*)[14]. AdipoR1 and AdipoR2 are predicted to contain a seven-transmembrane (7TM) domain[14], with an internal amino terminus and an external C terminus, which is the opposite configuration to G-protein-coupled receptors (GPCRs). Therefore, AdipoR1 and AdipoR2 are thought to be structurally and functionally distinct from GPCRs[15]. AdipoR1 and AdipoR2 serve as the major receptors for adiponectin *in vivo*, with AdipoR1 activating the AMPK pathways and AdipoR2 the PPAR-α pathways such as increased expression of uncoupling protein 2 (UCP2)[16]. Thereby, they regulate glucose and lipid metabolism, inflammation and oxidative stress *in vivo*. Recently, the small-molecule AdipoR agonist AdipoRon was shown to ameliorate diabetes and increase exercise endurance, and at the same time prolong the shortened lifespan in obesity[17]. It should also be noted that adiponectin receptors are conserved in evolution from mammals to plants and yeasts (http://www.ncbi.nlm.nih.gov/guide/proteins/)[18], strongly suggesting that they have essential biological roles.

It is extremely difficult to crystallize GPCRs, owing to their conformational complexity. By achieving technical breakthroughs, the crystal structures of the human β2 adrenoceptor (β2AR) were reported[19–21]. First, the conformational complexity of β2AR was controlled with high-affinity ligands (nanomolar dissociation constants), agonists and inverse agonists, to fix β2AR in the active and inactive forms, respectively[20–23]. Second, the crystallization was performed with antibody fragments and/or a protein fusion, in the lipidic mesophase. These technical advancements enabled the structure determination of many other GPCRs, and an understanding of their ligand specificities[24]. Furthermore, the first crystal structure of the active-state complex of an agonist-occupied β2AR with a nucleotide-free Gs heterotrimer was reported[25]. Thus, the β2AR structures greatly promoted the fields of GPCR research and drug development[26,27].

In contrast to the GPCRs, no information is available about the conformational states of AdipoR1 and AdipoR2 with respect to transmembrane signalling. Although the AdipoR agonist AdipoRon was successfully developed[17], further refinement of the AdipoR agonists to achieve nanomolar dissociation constants is still underway. The structural information about AdipoR1 and/or AdipoR2, if available, would be very important for understanding the AdipoR signalling mechanisms, and for developing and optimizing AdipoR agonists.

We optimized the properties of human AdipoR1 and AdipoR2 by deleting their N-terminal tails, and then used the Fv fragment of an anti-AdipoR monoclonal antibody and the lipidic mesophase for crystallization[28]. In this study, we successfully determined the crystal structures of human AdipoR1 and AdipoR2 at 2.9 and 2.4 Å resolution,

Figure 1 | Overall structures of AdipoR1 and AdipoR2. a, The 2.9 Å resolution structure of AdipoR1. **b**, The 2.4 Å resolution structure of AdipoR2. The structures were determined for their complexes with an Fv fragment, but the Fv fragments are omitted here for clarity. The structures are viewed from the extracellular side (left) and parallel to the membrane (right). The NTR, helix 0, transmembrane helices I–VII and the CTR of AdipoR1 (**a**) and AdipoR2 (**b**) are indicated.

respectively. The structures revealed their novel structural and functional properties, including the 7TM architecture, the zinc-binding site, and a putative adiponectin-binding surface, which are completely distinct from those of GPCRs, thus highlighting the uniqueness of the adiponectin receptors. This study should open new avenues towards the determination of an unprecedented model of signal transduction and the development and optimization of AdipoR agonists.

AdipoR1 and AdipoR2 in complexes with an Fv fragment

The N-terminally truncated constructs of human AdipoR1 and AdipoR2 (residues 89–375 and 100–386, respectively) exhibited better expression and purification properties than the full-length proteins[28]. These N-terminally truncated AdipoR1 and AdipoR2 displayed the same extents of adiponectin-stimulated AMPK phosphorylation[9–11] (Extended Data Fig. 1a) and *UCP2* upregulation[12,13] (Extended Data Fig. 1b), respectively, as those of the full-length proteins. Therefore, the N-terminally truncated AdipoR proteins were crystallized with the Fv fragment of a monoclonal antibody that recognizes a conformational epitope of both AdipoR1 and AdipoR2 in a cholesterol-doped monoolein lipidic mesophase[28]. Thus, we determined the crystal structures of AdipoR1 (Fig. 1a and Extended Data Fig. 1c–e) and AdipoR2 (Fig. 1b and Extended Data Fig. 1f–h) at 2.9 and 2.4 Å resolution, respectively. Data collection and refinement statistics are provided in Extended Data Table 1.

Structures of AdipoR1 and AdipoR2

The structure of AdipoR1 (residues 89–375) (Fig. 1a) contains the N-terminal intracellular region (residues 89–120; NTR), a short intracellular helix (residues 121–129; helix 0), the 7TM domain (residues

134–364), and the C-terminal extracellular region (residues 365–375; CTR) (Fig. 2). The Fv fragment is bound to the NTR (Extended Data Fig. 2a). The seven transmembrane helices (I–VII) are formed by residues 135–157, 169–192, 198–227, 232–252, 264–288, 305–319 and 336–364, respectively, and are connected by three intracellular loops (ICL1–3) and three extracellular loops (ECL1–3). ECL3 has a short α helix (residues 291–295; the ECL helix) in its centre, while ICL3 has another short α helix (residues 322–325; the ICL helix) just after helix VI. All of the residues are structurally ordered, except for residues 159–160 in ECL1, residues 298–299 in ECL3, and residues 374–375 in the CTR.

The seven transmembrane helices are bundled and arranged circularly in a clockwise manner, from helix I to VII, as viewed from the outside of the cell (Fig. 1a). The structure of AdipoR2 is quite similar to that of AdipoR1 (Fig. 1 and Extended Data Fig. 2a–c). The root mean squared deviation (r.m.s.d.) value for the main-chain Cα atoms between the AdipoR1 and AdipoR2 structures is as small as 0.56 Å.

The DALI search[29] indicated that the AdipoR1 and AdipoR2 structures share no similarity with other structures in the Protein Data Bank (PDB). The C-terminus-out topology of the 7TM domain of AdipoR1/AdipoR2, relative to the plasma membrane, is opposite to the N-terminus-out topology of the conventional 7TM proteins, such as GPCRs[30] and microbial rhodopsins[31]. Furthermore, the conformational characteristics, such as the proline-induced kink[19–21], of the transmembrane helices of GPCRs in classes A, B and C (refs 20, 21, 32, 33) are not observed for those of AdipoR1/AdipoR2 (Extended Data Fig. 3). In the AdipoR1/AdipoR2 structures, the transmembrane helices are not kinked, while helix V is slightly curved owing to three Gly residues (Fig. 2). Consequently, we concluded that the AdipoR1 and AdipoR2 structures are novel.

The zinc-binding sites of AdipoR1 and AdipoR2

Remarkably, we found a zinc ion bound within the 7TM domain in the AdipoR1 and AdipoR2 structures (Fig. 3a), by X-ray absorption spectroscopy (data not shown) and the anomalous difference Fourier map (Fig. 3b). The zinc-binding site is located in the intracellular layer of the membrane. The zinc ion is coordinated by three His residues, His 191 in helix II and His 337 and His 341 in helix VII of AdipoR1, and His 202 in helix II and His 348 and His 352 in helix VII of AdipoR2, at zinc–nitrogen distances of 2.1–2.6 Å (Fig. 3c, d). The zinc ion is thus located approximately 4 Å deep from the inner surface of the plasma membrane (Fig. 3a). Furthermore, a water molecule is observed between the zinc ion and the side-chain carboxyl group of Asp 219 in helix III of AdipoR2. Thus, the zinc ion has a tetrahedral coordination (Fig. 3d). The zinc ion binds helices II, III and VII together (Fig. 3c, d), and probably stabilizes the structure of the subdomain consisting of helices I, II, III and VII (Extended Data Fig. 2). The three His and Asp (3×His+Asp) residues of AdipoR1 and AdipoR2 are strictly conserved in the homologues from mammals to plants and bacteria (Extended Data Fig. 4a, b).
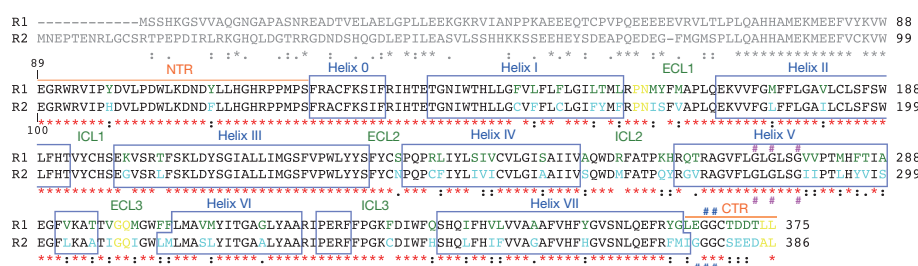
Figure 2 | Sequence alignment of human AdipoR1 and AdipoR2. Amino acid residues that are not conserved between these receptors are shown in green (AdipoR1) and cyan (AdipoR2). The deleted residues in the constructs and the disordered residues in the crystal structures are shown in grey and yellow, respectively. The helices in the crystal structures are surrounded by blue squares. The identical and similar residues between the two proteins are indicated with red asterisks and black colons, respectively. The characteristic Gly residues in helix V and in the CTR are indicated with red and blue number signs, respectively.

Figure 3 | The zinc-binding sites of AdipoR1 and AdipoR2. a, The position of the zinc ion (magenta sphere) in the AdipoR2 structure. b, The anomalous difference maps of AdipoR2 calculated from the peak data set (red, 1.28 Å) and the low remote data set (blue, 1.288 Å), at a resolution of 3.0 Å and contoured at 3.0σ. Helix I has been omitted for clarity. c, d, Coordination of the zinc ion (magenta sphere) by three His residues of AdipoR1 (c) and AdipoR2 (d), viewed from the cytoplasmic side. A water molecule (pink sphere) is also coordinated to the zinc ion, and is fixed by Asp 219 in AdipoR2. e, Phosphorylation and amounts of AMPK in HEK293 cells transfected with AdipoR1 (residues 89–375) or its mutants (see text), treated for 5 min with adiponectin (15 μg ml^{-1}). f, UCP2 mRNA levels in HEK293 cells transfected with AdipoR2 (residues 100–386) or its mutants (see text), treated for 18 h with adiponectin (3 μg ml^{-1}). The ratio of UCP2 mRNA to the housekeeping gene PPIA (cyclophilin A) was used for normalization. All values are mean ± s.e.m. n = 3–4, three independent experiments.*$P < 0.05$, **$P < 0.01$ compared to control cells or as indicated (see Methods for statistical tests used). Ad, adiponectin.

We mutated the zinc-coordinated 3× His + Asp residues of AdipoR1 (residues 89–375) (Fig. 3e). As compared with the parent AdipoR1 molecule (89–375), the adiponectin-stimulated AMPK phosphorylation was reduced by the triple mutant His191Ala/His337Ala/His341Ala (3Ala) and more seriously by the quadruple mutant His191Ala/Asp208Ala/His337Ala/His341Ala (4Ala), while none of the single His191Ala, Asp208Ala, His337Ala and His341Ala mutations affected it (Fig. 3e and Extended Data Fig. 4c). Therefore, the results suggested that zinc binding is not directly required for the adiponectin-stimulated AMPK phosphorylation, but exerts a putative structure-stabilizing effect.

By contrast, the adiponectin-stimulated UCP2 upregulation by AdipoR2 was markedly reduced by each of the single mutations Asp219Ala and His348Ala, and nearly completely eliminated by the triple mutation His202Ala/His348Ala/His352Ala (3Ala) and the

quadruple mutation His202Ala/Asp219Ala/His348Ala/His352Ala (4Ala) of AdipoR2 (residues 1–386 and 100–386), as compared with the wild-type AdipoR2 (Fig. 3f and Extended Data Fig. 4d, e). Correspondingly, the single mutations His202Ala and His352Ala of AdipoR2 (residues 100–386) did not decrease the amount of bound zinc ion, whereas the single mutations Asp219Ala and His348Ala decreased it moderately, and the multiple mutations 3Ala and 4Ala reduced it markedly (data not shown). These results suggested that the zinc ion is directly involved in the adiponectin-stimulated UCP2 upregulation in the case of AdipoR2, in addition to structural stabilization.

An attractive hypothesis is that AdipoR2 has zinc-ion-dependent hydrolytic activity, and uses the water molecule fixed between the zinc ion and the side-chain carboxyl group of Asp219 of AdipoR2 for the nucleophilic attack on the carbonyl carbon atom of substrates. Free fatty acid might be produced from lipid hydrolysis by the adiponectin-stimulated AdipoR2, and PPAR-α activation by the produced free fatty acid would increase the expression of the target genes, such as UCP2.

The zinc-binding structures in the transmembrane domains of AdipoR1 and AdipoR2 are novel. The only previously reported membrane protein structure with a zinc ion within the transmembrane domain is that of a site-2 protease family intramembrane metalloprotease[34]. The protease consists of six transmembrane segments, and the catalytic zinc ion is coordinated by two His residues and one Asp residue, and is approximately 14 Å deep from the inner surface of the plasma membrane. Therefore, the site-2 protease and AdipoR structures are not homologous. By contrast, some globular zinc enzyme structures share architectural similarity, in terms of the coordination of three His residues and a water molecule[35,36] (Extended Data Fig. 5). Although the transmembrane alkaline ceramidases share negligible sequence homology

Figure 4 | The large internal cavities in the AdipoR1 and AdipoR2 structures. a, b, The cavities of AdipoR1 (a) and AdipoR2 (b). Red arrows indicate the openings of the cavities. c, d, The extra electron density maps in the cavities of AdipoR1 (c) and AdipoR2 (d) contoured at 0.5σ and 1σ, respectively. The NTR (residues 89–119) is coloured orange. The openings of the cavities are bordered in red and pointed with arrows.

with AdipoR, three His residues and one Asp residue are conserved in these proteins. However, their crystal structures have not been solved. Therefore, we presently cannot completely exclude the possibility that the AdipoRs have ceramidase activity.

The large internal cavities of AdipoR1 and AdipoR2

In both the AdipoR1 and AdipoR2 structures, the seven transmembrane helices surround a large internal cavity, including the zinc-binding site (Fig. 4a, b). This large internal cavity is formed between the four- and three-helix subdomains (helices VII-I-II-III and IV-V-VI, respectively) of the 7TM domains of AdipoR1/2 (Extended Data Fig. 2). The cavities extend from the cytoplasmic surface to the middle of the outer lipid layer of the membrane (Fig. 4a, b), and contain unidentified extra electron densities, which are weaker than those of the protein (Fig. 4c, d). In the cavity of AdipoR2, the extra electron densities are observed along with helices III, V and VI (Fig. 4d). By contrast, in the cavity of AdipoR1, even weaker electron densities are observed on the cytoplasmic side of the cavity (Fig. 4c). These weak electron densities might be relevant to the substrates/products of the hypothesized hydrolytic activities of AdipoR1/AdipoR2.

The cavity has small openings between helices V and VI within the outer lipid layer and between helices IV and VI on the cytoplasmic side (Fig. 4a, b). Intriguingly, a much larger opening at helices III–VII would be uncovered on the cytoplasmic side, if the NTR was displaced from its present position (Extended Data Fig. 6). These openings might serve as the entrance/exit for the substrate/product of the hypothesized hydrolytic activity. Notably, the shorter constructs (residues 102–375 and 120–375) of AdipoR1 are also as active as the full-length AdipoR1 with respect to adiponectin-stimulated AMPK phosphorylation (Extended Data Fig. 1a), indicating that the NTR, which covers the large internal cavity, is not required for this activity.

The amino acid sequences of the ICL2 regions are significantly different between AdipoR1 and AdipoR2 (Fig. 2). In particular, AdipoR1 has a cluster of positively charged residues, Arg 257, Lys 262 and His 263, in the ICL2 region (Extended Data Fig. 7), unlike AdipoR2. Consequently, this structural difference in the cytoplasmic face may reflect the distinct signalling pathways downstream of these adiponectin receptors.

The extracellular faces of AdipoR1 and AdipoR2

The ECL1–3 and the CTR are exposed on the extracellular faces of AdipoR1 and AdipoR2. The three extracellular loops exhibit high conservation between AdipoR1 and AdipoR2 (Fig. 5a–d). Helices VII and III are longer than the others, and the C-terminal two turns of helix VII protrude from the extracellular face. The CTR, which follows helix VII, seems to be independent of the other extracellular structural elements, the ECL1–3 and helix VII (Fig. 1). The very C-terminal Leu 374–Leu 375 (AdipoR1) and Ala 385–Leu 386 (AdipoR2) residues are disordered, and the crystal packing fixed the tail conformations differently (Extended Data Fig. 8). Therefore, the entire CTRs of AdipoR1 and AdipoR2 are likely to be flexible and unstructured (Figs 1 and 2).

Adiponectin should bind to the extracellular face of the receptor, and the adiponectin-binding site seems to be shared by AdipoR1 and AdipoR2. A yeast two-hybrid analysis revealed that adiponectin interacts with a C-terminal fragment[37], which extends from the middle of helix VI to the very C terminus in the present structures. On the other hand, in this study, the CTR deletion after residue 366 or 370 of AdipoR1 did not affect the adiponectin-stimulated AMPK phosphorylation via AdipoR1 (Fig. 5e), indicating that the flexible CTR is not necessarily required for AMPK phosphorylation by adiponectin. By contrast, the longer deletion of the C-terminal thirteen residues up to Tyr 363 and Gly 364, the last two residues of helix VII, reduced the adiponectin-stimulated AMPK phosphorylation via AdipoR1 (residues 1–362; Fig. 5e and Extended Data Fig. 9a), indicating that the protruding C-terminal turn of helix VII may be involved in adiponectin signalling. Furthermore, the extracellular loop residues conserved between AdipoR1 and AdipoR2 were mutated to Gly/Ser (Fig. 5f, g): the three-loop mutation
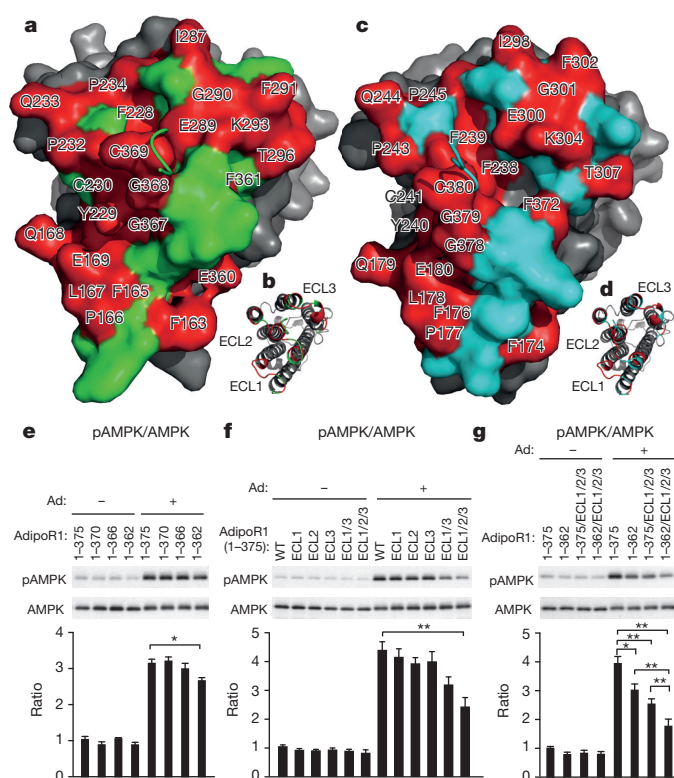
Figure 5 | The extracellular faces of AdipoR1 and AdipoR2. a–d, The extracellular faces of AdipoR1 (**a, b**) and AdipoR2 (**c, d**). AdipoR1 and AdipoR2 are shown by surface (**a, c**) and cartoon (**b, d**) representations. The residues conserved between AdipoR1 and AdipoR2 are shown in red and labelled in black. The AdipoR1- and AdipoR2-specific residues are shown in green and cyan, respectively. The 7TM domains of AdipoR1 and AdipoR2 are shown in grey. The CTR residues 370-Thr-Asp-Asp-372 and 381-Glu-Glu-Asp-383 of AdipoR1 and AdipoR2, respectively, were removed for clarity. **e–g,** Phosphorylation and amounts of AMPK in HEK293 cells transfected with full-length AdipoR1 (residues 1–375) or a variety of mutants of AdipoR1, treated for 5 min with adiponectin (15 μg ml⁻¹). All values are mean ± s.e.m. $n = 3$–4, three independent experiments.*$P < 0.05$, **$P < 0.01$ compared to control cells or as indicated (see Methods). ECL1, MYFMAPL (residues 161–167) changed to SGSSGGS; ECL2, YCS (residues 229–231) changed to GGG; ECL3, FVKATTV (residues 291–297) changed to SSSGGGS; ECL1/3, ECL1 and ECL3; ECL1/2/3, ECL1, ECL2 and ECL3. WT, wild type.

(ECL1/2/3) combined with the C-terminal 13-residue deletion (1–362) remarkably decreased adiponectin-stimulated AMPK phosphorylation via AdipoR1 (Fig. 5g and Extended Data Fig. 9b). The other mutants with fewer Gly/Ser mutations (ECL1, ECL2, ECL3 and ECL1/3) or with no C-terminal deletion showed correspondingly smaller decreased (Fig. 5f, g and Extended Data Fig. 9a, b). These data raised the possibility that AdipoR1 may recognize adiponectin by the extensive use of its extracellular face, including the three extracellular loops and the C-terminal turns of helix VII.

Conclusions

The structural and functional characteristics of AdipoR1 and AdipoR2 revealed by this study are completely different from those of GPCRs, and therefore the AdipoRs represent an entirely new class of receptor. The present crystal structures are expected to provide a strong basis for the development and optimization of adiponectin receptor agonists, such as AdipoRon[17], as well as for understanding the roles and mechanisms of the AdipoR1/AdipoR2 homologues from animals and plants in putative signalling, such as in defence systems and lipid metabolism (Extended Data Fig. 4a, b).

1. Scherer, P. E., Williams, S., Fogliano, M., Baldini, G. & Lodish, H. F. A novel serum protein similar to C1q, produced exclusively in adipocytes. *J. Biol. Chem.* **270**, 26746–26749 (1995).
2. Hu, E., Liang, P. & Spiegelman, B. M. AdipoQ is a novel adipose-specific gene dysregulated in obesity. *J. Biol. Chem.* **271**, 10697–10703 (1996).
3. Maeda, K. *et al.* cDNA cloning and expression of a novel adipose specific collagen-like factor, apM1 (AdiPose Most abundant Gene transcript 1). *Biochem. Biophys. Res. Commun.* **221**, 286–289 (1996).
4. Nakano, Y., Tobe, T., Choi-Miura, N. H., Mazda, T. & Tomita, M. Isolation and characterization of GBP28, a novel gelatin-binding protein purified from human plasma. *J. Biochem.* **120**, 803–812 (1996).
5. Hotta, K. *et al.* Plasma concentrations of a novel, adipose-specific protein, adiponectin, in type 2 diabetic patients. *Arterioscler. Thromb. Vasc. Biol.* **20**, 1595–1599 (2000).
6. Yamauchi, T. *et al.* The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nature Med.* **7**, 941–946 (2001).
7. Berg, A. H., Combs, T. P., Du, X., Brownlee, M. & Scherer, P. E. The adipocyte-secreted protein Acrp30 enhances hepatic insulin action. *Nature Med.* **7**, 947–953 (2001).
8. Fruebis, J. *et al.* Proteolytic cleavage product of 30-kDa adipocyte complement-related protein increases fatty acid oxidation in muscle and causes weight loss in mice. *Proc. Natl Acad. Sci. USA* **98**, 2005–2010 (2001).
9. Yamauchi, T. *et al.* Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase. *Nature Med.* **8**, 1288–1295 (2002).
10. Tomas, E. *et al.* Enhanced muscle fat oxidation and glucose transport by ACRP30 globular domain: acetyl-CoA carboxylase inhibition and AMP-activated protein kinase activation. *Proc. Natl Acad. Sci. USA* **99**, 16309–16313 (2002).
11. Kahn, B. B., Alquier, T., Carling, D. & Hardie, D. G. AMP-activated protein kinase: ancient energy gauge provides clues to modern understanding of metabolism. *Cell Metab.* **1**, 15–25 (2005).
12. Kersten, S., Desvergne, B. & Wahli, W. Roles of PPARs in health and disease. *Nature* **405**, 421–424 (2000).
13. Yamauchi, T. *et al.* Globular adiponectin protected ob/ob mice from diabetes and ApoE-deficient mice from atherosclerosis. *J. Biol. Chem.* **278**, 2461–2468 (2003).
14. Yamauchi, T. *et al.* Cloning of adiponectin receptors that mediate antidiabetic metabolic effects. *Nature* **423**, 762–769 (2003).
15. Wess, J. G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of G-protein recognition. *FASEB J.* **11**, 346–354 (1997).
16. Yamauchi, T. *et al.* Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions. *Nature Med.* **13**, 332–339 (2007).
17. Okada-Iwabu, M. *et al.* A small-molecule AdipoR agonist for type 2 diabetes and short life in obesity. *Nature* **503**, 493–499 (2013).
18. Lyons, T. J. *et al.* Metalloregulation of yeast membrane steroid receptor homologs. *Proc. Natl Acad. Sci. USA* **101**, 5506–5511 (2004).
19. Rasmussen, S. G. *et al.* Crystal structure of the human β2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387 (2007).
20. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science* **318**, 1258–1265 (2007).
21. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β2-adrenergic receptor function. *Science* **318**, 1266–1273 (2007).
22. Rasmussen, S. G. *et al.* Structure of a nanobody-stabilized active state of the β2 adrenoceptor. *Nature* **469**, 175–180 (2011).
23. Rosenbaum, D. M. *et al.* Structure and function of an irreversible agonist-β2 adrenoceptor complex. *Nature* **469**, 236–240 (2011).
24. Venkatakrishnan, A. J. *et al.* Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185–194 (2013).
25. Rasmussen, S. G. *et al.* Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011).
26. Shimamura, T. *et al.* Structure of the human histamine H1 receptor complex with doxepin. *Nature* **475**, 65–70 (2011).
27. de Graaf, C. *et al.* Crystal structure-based virtual screening for fragment-like ligands of the human histamine H1 receptor. *J. Med. Chem.* **54**, 8195–8206 (2011).
28. Tanabe, H. *et al.* Expression, purification, crystallization, and preliminary X-ray crystallographic studies of the human adiponectin receptors, AdipoR1 and AdipoR2. *J. Struct. Funct. Genomics* **16**, 11–23 (2015).
29. Holm, L. & Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 (2010).
30. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745 (2000).
31. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science* **277**, 1676–1681 (1997).
32. Siu, F. Y. *et al.* Structure of the human glucagon class B G-protein-coupled receptor. *Nature* **499**, 444–449 (2013).
33. Wu, H. *et al.* Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science* **344**, 58–64 (2014).
34. Feng, L. *et al.* Structure of a site-2 protease family intramembrane metalloprotease. *Science* **318**, 1608–1612 (2007).
35. Bode, W., Gomis-Ruth, F. X., Huber, R., Zwilling, R. & Stocker, W. Structure of astacin and implications for activation of astacins and zinc-ligation of collagenases. *Nature* **358**, 164–167 (1992).
36. Eriksson, A. E., Jones, T. A. & Liljas, A. Refined structure of human carbonic anhydrase II at 2.0 Å resolution. *Proteins* **4**, 274–282 (1988).
37. Mao, X. *et al.* APPL1 binds to adiponectin receptors and mediates adiponectin signalling and function. *Nature Cell Biol.* **8**, 516–523 (2006).

Author Contributions H.T., T.K.-S., M.S., M.O.-I., M. Iwabu, T.Y., T.K. and S.Y. designed the research. H.T., K.M., M. Ikeda, M.W. and T.T. performed protein expression, purification and analyses of AdipoR1 and AdipoR2, while N.O. designed and constructed the expression plasmids. M.H. provided the lipidic mesophase crystallization techniques, and H.T. and K.M. performed the crystallization of the receptors. Y.F., H.T., Y.N. and T. Hosaka performed the X-ray diffraction data collection and the structural analysis. K.H., Y.K. and M.Y. designed the microcrystal data collection strategy, using BL32XU at SPring-8. H.T. prepared the AdipoR1 and AdipoR2 immunogens, and H.T., K.M., S.O., T. Hino, T.M. and S.I. produced the anti-AdipoR1 monoclonal antibody. M.O.-I. and M. Iwabu assayed the activities of the mutants. H.T., T.K.-S., M.S., M.O.-I., M. Iwabu, T.Y., T.K. and S.Y. wrote the manuscript. All authors commented on the manuscript.

Author Information The atomic coordinates and structure factors for the AdipoR1–Fv and AdipoR2–Fv structures have been deposited in the Protein Data Bank under the accession codes 3WXV and 3WXW, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.Y. (yokoyama@riken.jp), T.K. (kadowaki-3im@h.u-tokyo.ac.jp), or T.Y. (tyamau-tky@umin.net).

METHODS

Preparation of the AdipoR1–Fv and AdipoR2–Fv crystals. The human AdipoR1 and AdipoR2 proteins and the Fv fragment of an anti-AdipoR1 monoclonal antibody were prepared as described[28]. In brief, human AdipoR1 and AdipoR2 (residues 89–375 and 100–386, respectively) were expressed in High Five insect cells. The proteins were purified by Flag antibody affinity chromatography followed by anion exchange chromatography, metal ion affinity chromatography after cleaving the N-terminal Flag tag by His-tagged tobacco etch virus (TEV) protease, and size-exclusion chromatography. The Fv fragment was cloned from hybridoma cells. The Fv fragment was synthesized by the *Escherichia coli* cell-free protein synthesis method, and purified by Ni-affinity chromatography followed by size-exclusion chromatography. The purified AdipoR1 and AdipoR2 proteins were mixed with the Fv fragment, and the AdipoR1–Fv and AdipoR2–Fv complexes were purified by size-exclusion chromatography, and crystallized by the lipidic mesophase method[38].

X-ray data collection. Data collection was performed on beamline BL32XU at SPring-8, using an MX225HE CCD detector[39–41]. X-ray diffraction data were collected at 100 K by the helical scan method, with a beam size of 1×10 μm (horizontal $\times$ vertical) using $1°$ oscillation. The AdipoR1 and AdipoR2 crystals diffracted up to 2.8 Å and 2.2 Å resolution, respectively[28]. Data collection from the AdipoR1 crystals was limited to 10–30 images per crystal, owing to radiation damage in the microcrystals, and data from five crystals were merged to complete the data set. For AdipoR2, diffraction data were collected from a single crystal. The data from the AdipoR1 crystals and the AdipoR2 crystal were indexed, scaled and merged with the HKL2000 program suite[42] and the XDS package[43], respectively. The data collection statistics are shown in Extended Data Table 1. The AdipoR1 crystals belonged to the space group $C222_1$, with unit cell parameters $a = 92.3$, $b = 194.1$, $c = 74.3$ Å, and the AdipoR2 crystal belonged to the space group $P2_12_12$, with unit cell parameters $a = 74.6$, $b = 108.6$, $c = 101.0$ Å.

Structure solution and refinement. The initial phases for the AdipoR2–Fv complex were obtained by molecular replacement, using Fv (the V_H and V_L fragments from PDB accessions 1E6J and 1FDL, respectively) in Phaser[44] as a search model. The resulting phases were improved by density modification using the program RESOLVE[45], and thereby the electron density map around the helix bundle region of AdipoR2 became clearly visible. The initial model (all of Fv and about 80% of AdipoR2) was automatically built using the program AutoBuild[46], and the rest of the model (the loops connecting the transmembrane helices) was built manually using COOT[47]. Refinement was performed with phenix.refine[48], and the refined coordinates were rebuilt with COOT. The structure of the AdipoR2–Fv complex was refined with final R_{work}/R_{free} values of 0.25/0.29. The structure of the AdipoR1–Fv complex was determined by molecular replacement, using that of the AdipoR2–Fv complex as a search model, and was refined with the secondary structure restraints in phenix.refine. Refinement of the AdipoR1–Fv complex was performed similarly to that of the AdipoR2–Fv complex. The structure of the AdipoR1–Fv complex was refined with final R_{work}/R_{free} values of 0.24/0.30. Ramachandran statistics were analysed with MolProbity[49]. In the AdipoR1–Fv complex structure, 95.8% of residues were in favoured regions and 4.2% of residues were in allowed regions. In the AdipoR2–Fv complex structure, 95.6% of residues were in favoured regions and 4.4% of residues were in allowed regions. Each of the final models of the AdipoR1–Fv and AdipoR2–Fv complexes includes 281 residues of the receptor, 119 residues of V_H, and 107 residues of V_L. The data collection and refinement statistics are summarized in Extended Data Table 1. Structural illustrations were generated using PyMol[50].

Cell culture. HEK293T cells (ATCC) were cultured in DMEM supplemented with 10% (v/v) FBS. Cells were transfected using Lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. The cDNAs encoding the *ADIPOR* mutants were introduced into the pOriP vector, for the expression of proteins tagged with a Flag epitope.

Generation of recombinant adiponectin. Recombinant mouse full-length adiponectin was generated as previously described[6,9,13,14,17,51]. The expression of His-tagged adiponectin was induced by the addition of isopropyl β-D-1-thiogalactopyranoside to the growth medium. Bacterial extracts were prepared using standard methods, and the fusion proteins were purified by elution through a nickel-ion agarose column.

Western blot analysis and measurement of AMPK activities. Phosphorylation and protein levels of αAMPK were determined as described[52–55]. Western blot analyses were performed with anti-phosphorylated-AMPK (Cell Signaling Technology 2535) and anti-αAMPK (Cell Signaling Technology 2532) antibodies. Protein levels of AdipoR were analysed by western blotting, using an anti-Flag antibody (Sigma-Aldrich F1804).

Real-time PCR. Real-time PCR was performed according to the method described previously[16,51]. Total RNA was prepared from cells with Trizol (Invitrogen), according to the manufacturer's instructions. We used the real-time PCR method to quantify the mRNAs[14], with slight modifications.

Statistics. Results are expressed as mean ± s.e.m. Differences between two groups were assessed using unpaired two-tailed *t*-tests. Data involving more than two groups were assessed by analysis of variance (ANOVA) followed by post-hoc comparisons. No statistical methods were used to predetermine sample size.

38. Hato, M., Hosaka, T., Tanabe, H., Kitsunai, T. & Yokoyama, S. A new manual dispensing system for *in meso* membrane protein crystallization with using a stepping motor-based dispenser. *J. Struct. Funct. Genomics* **15**, 165–171 (2014).
39. Hirata, K. *et al.* Achievement of protein micro-crystallography at SPring-8 beamline BL32XU. *J. Phys. Conf. Ser.* **425**, 012002 (2013).
40. Murakami, I. *et al.* Tumor volume and lymphovascular space invasion as a prognostic factor in early invasive adenocarcinoma of the cervix. *J. Gynecol. Oncol.* **23**, 153–158 (2012).
41. Ueno, G., Kanda, H., Kumasaka, T. & Yamamoto, M. Beamline Scheduling Software: administration software for automatic operation of the RIKEN structural genomics beamlines at SPring-8. *J. Synchrotron Radiat.* **12**, 380–384 (2005).
42. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
43. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
44. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
45. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56**, 965–972 (2000).
46. Terwilliger, T. C. Automated side-chain model building and sequence assignment by template matching. *Acta Crystallogr. D* **59**, 45–49 (2003).
47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
48. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
49. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
50. DeLano, W. L. The PyMOL molecular graphics system. http://www.pymol.org (DeLano Scientific, 2002).
51. Iwabu, M. *et al.* Adiponectin and AdipoR1 regulate PGC-1α and mitochondria by Ca^{2+} and AMPK/SIRT1. *Nature* **464**, 1313–1319 (2010).
52. Minokoshi, Y. *et al.* Leptin stimulates fatty-acid oxidation by activating AMP-activated protein kinase. *Nature* **415**, 339–343 (2002).
53. Tsao, T. S., Murrey, H. E., Hug, C., Lee, D. H. & Lodish, H. F. Oligomerization state-dependent activation of NF-κB signaling pathway by adipocyte complement-related protein of 30 kDa (Acrp30). *J. Biol. Chem.* **277**, 29359–29362 (2002).
54. Woods, A., Salt, I., Scott, J., Hardie, D. G. & Carling, D. The α1 and α2 isoforms of the AMP-activated protein kinase have similar activities in rat liver but exhibit differences in substrate specificity *in vitro*. *FEBS Lett.* **397**, 347–351 (1996).
55. Hayashi, T. *et al.* Metabolic stress and altered glucose transport. Activation of AMP-activated protein kinase as a unifying coupling mechanism. *Diabetes* **49**, 527–531 (2000).
56. Hanson, M. A. *et al.* Crystal structure of a lipid G protein-coupled receptor. *Science* **335**, 851–855 (2012).
57. Lebon, G. *et al.* Agonist-bound adenosine A_{2A} receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011).
58. Vogeley, L. *et al.* Anabaena sensory rhodopsin: a photochromic color sensor at 2.0 Å. *Science* **306**, 1390–1393 (2004).

ARTICLE

Two disparate ligand–binding sites in the human P2Y₁ receptor

Dandan Zhang[1], Zhan-Guo Gao[2], Kaihua Zhang[1], Evgeny Kiselev[2], Steven Crane[2], Jiang Wang[1], Silvia Paoletta[2], Cuiying Yi[1], Limin Ma[1], Wenru Zhang[1], Gye Won Han[3], Hong Liu[1], Vadim Cherezov[3], Vsevolod Katritch[4], Hualiang Jiang[5], Raymond C. Stevens[3,4,6], Kenneth A. Jacobson[2], Qiang Zhao[1] & Beili Wu[1]

In response to adenosine 5′-diphosphate, the P2Y₁ receptor (P2Y₁R) facilitates platelet aggregation, and thus serves as an important antithrombotic drug target. Here we report the crystal structures of the human P2Y₁R in complex with a nucleotide antagonist MRS2500 at 2.7 Å resolution, and with a non-nucleotide antagonist BPTU at 2.2 Å resolution. The structures reveal two distinct ligand-binding sites, providing atomic details of P2Y₁R's unique ligand-binding modes. MRS2500 recognizes a binding site within the seven transmembrane bundle of P2Y₁R, which is different in shape and location from the nucleotide binding site in the previously determined structure of P2Y₁₂R, representative of another P2YR subfamily. BPTU binds to an allosteric pocket on the external receptor interface with the lipid bilayer, making it the first structurally characterized selective G-protein-coupled receptor (GPCR) ligand located entirely outside of the helical bundle. These high-resolution insights into P2Y₁R should enable discovery of new orthosteric and allosteric antithrombotic drugs with reduced adverse effects.

Human purinergic GPCRs are divided into two subfamilies, G_q-coupled P2Y₁R-like receptors and G_i-coupled P2Y₁₂R-like receptors[1]. Both P2Y₁R and P2Y₁₂R are activated by adenosine 5′-diphosphate (ADP) to induce platelet activation, which plays a pivotal role in thrombosis formation[2,3]. The blockade of either receptor significantly decreases ADP-induced platelet aggregation[4]. Although most of the available antithrombotic drugs act on P2Y₁₂R, P2Y₁R has been suggested as a new promising target, which may offer a safety advantage over P2Y₁₂R inhibitors in terms of reduced bleeding liabilities[2]. Besides platelet aggregation, P2Y₁R is also involved in many other physiological processes, such as vascular inflammation, and Ca^{2+} wave propagation and activation of extracellular signal-regulated kinase in astrocytes[5–7].

(1′R,2′S,4′S,5′S)-4-(2-Iodo-6-methylaminopurin-9-yl)-1-[(phosphato) methyl]-2(phosphato)bicyclo[3.1.0]-hexane (MRS2500) is a potent P2Y₁R antagonist that completely blocks ADP-induced platelet aggregation and effectively reduces arterial thrombosis with only a moderate prolongation of the bleeding time, which makes it an attractive candidate as an antithrombotic agent[8,9]. Another ligand, 1-(2-(2-(tert-butyl) phenoxy)pyridin-3-yl)-3-(4-(trifluoromethoxy)phenyl)urea (BPTU), was recently discovered by Bristol-Myers Squibb as a novel P2Y₁R antagonist that substantially reduces platelet aggregation with a minimal effect on bleeding, for the treatment of thrombosis[10]. To understand how these antithrombotic ligands recognize their purinoceptor target and to enable new drug discovery, we solved X-ray crystal structures of the human P2Y₁R receptor bound to MRS2500 and BPTU (Extended Data Table 1), and performed a structure-guided mutagenesis study (Extended Data Table 2).

Overall architecture of P2Y₁R

The P2Y₁R structures share a canonical seven-transmembrane helical bundle architecture with other known GPCR structures (Fig. 1a, b and Extended Data Fig. 1). There are two disulfide bonds, connecting the N terminus to helix VII, and helix III to the second extracellular loop (ECL2), stabilizing the conformations of the ECLs. The ECL2 of P2Y₁R exhibits a hairpin structure, which was previously observed in all the known peptide-bound GPCR structures, such as PAR1, NTSR1, chemokine and opioid receptors[11–15]. The conserved D[E]R^{3.50}Y motif in class A GPCR family is replaced by an HR^{3.50}Y motif in P2Y₁R, making the crystal structure of P2Y₁R the first GPCR structure with a basic histidine residue at position 3.49 (Ballesteros–Weinstein nomenclature[16]) (Extended Data Fig. 2a). Distinct from many other class A GPCRs, which contain a salt bridge between D[E]^{3.49} and R^{3.50}, H148^{3.49} in P2Y₁R repels R149^{3.50}, resulting in a more extended side chain conformation of this residue. Consequently, R149^{3.50} forms a hydrogen bond with the main chain of A327^{7.56} at the intracellular tip of helix VII and stabilizes the C terminus in a different conformation compared to many other known class A GPCR structures. Like PAR1 and some other GPCRs, the P2Y₁R structures lack helix VIII, and the C-terminal region beyond R338 appears disordered.

The two P2Y₁R structures are similar (Cα root mean squared deviation (r.m.s.d.) within the entire receptor is 0.9 Å), except for subtle differences at the extracellular ends of helices I and II. Additionally, in the P2Y₁R–MRS2500 structure, a salt bridge between R195 in ECL2 and MRS2500 shifts the β-hairpin tip of ECL2 by 2.6 Å towards the central axis of the helical bundle compared to the P2Y₁R–BPTU structure (Extended Data Fig. 2b). Compared with the recently solved structure of P2Y₁₂R[17,18] belonging to a separate G_i-coupled P2YR subfamily, P2Y₁R is structurally distinct from either the agonist-bound or the antagonist-bound P2Y₁₂R, with Cα r.m.s.d. within the helical bundle of 2.2 and 2.6 Å, respectively (Fig. 1c, d). The extracellular tip of P2Y₁R's helix VI has a position intermediate between the agonist and antagonist-bound P2Y₁₂R structures; whereas helix VII in P2Y₁R is in a relatively similar conformation to the antagonist-bound P2Y₁R structure. Unlike P2Y₁₂R, P2Y₁R has a highly conserved in class A GPCR residue P229^{5.50},

[1]CAS Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Pudong, Shanghai 201203, China. [2]Molecular Recognition Section, Laboratory of Bioorganic Chemistry, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [3]Bridge Institute, Department of Chemistry, University of Southern California, Los Angeles, California 90089, USA. [4]Bridge Institute, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA. [5]Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Pudong, Shanghai 201203, China. [6]iHuman Institute, ShanghaiTech University, 99 Haike Road, Pudong, Shanghai 201203, China.
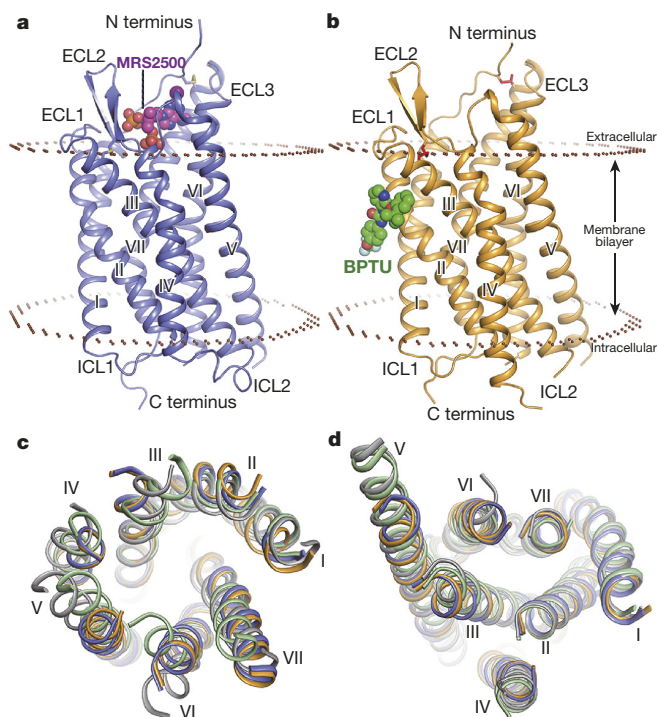
Figure 1 | Structures of the P2Y₁R–MRS2500 and P2Y₁R–BPTU complexes. a, b, Side view of the P2Y₁R–MRS2500 (**a**) and P2Y₁R–BPTU (**b**) structures. The receptor is shown in blue (**a**) and orange (**b**) cartoon representation. The ligands MRS2500 and BPTU are shown in sphere representation with magenta and green carbons, respectively. The disulfide bonds are shown as yellow (**a**) and red (**b**) sticks. The membrane boundaries (brown) are adapted from the OPM database (http://opm.phar.umich.edu/) with P2Y₁₂R (PDB ID: 4NTJ) as a model. **c, d**, Structural comparison of the helical bundles between P2Y₁R and P2Y₁₂R. **c**, Top view of the extracellular side. **d**, Bottom view of the intracellular side. The receptors are in cartoon representation. The P2Y₁R–MRS2500, P2Y₁R–BPTU, P2Y₁₂R–AZD1283 (PDB ID: 4NTJ) and P2Y₁₂R–2MeSADP (PDB ID: 4PXZ) structures are coloured blue, orange, grey and green, respectively.

which leads to a helical kink and displaces the extracellular end of helix V by over 4 Å compared to the P2Y₁₂R structures. Another substantial difference between the helical bundles of the two purinergic receptors is that the extracellular end of helix III shifts away from the axis of the seven-transmembrane helical bundle by over 5 Å in P2Y₁R compared to the P2Y₁₂R structures. This apparent shift is probably due to the different conformations that the ECL2 adopts in P2Y₁R and P2Y₁₂R. The intracellular halves of the two P2Y receptors, however, are very similar with all helices overlaying each other relatively well.

Ligand-binding mode of P2Y₁R–MRS2500

In the P2Y₁R–MRS2500 structure, the ligand occupies a pocket defined by residues mainly from the N terminus, ECL2 and helices VI and VII (Fig. 2 and Extended Data Fig. 3a). The adenine ring of MRS2500 inserts into a binding crevice with R287[6.62] and L44 on either side, and its N⁶H and N⁷ are coordinated by two hydrogen bonds with the N283[6.58] side chain, similar to adenine recognition in A₂ₐ adenosine receptor (N[6.55]) and P2Y₁₂R (N[5.40]) but at a different location. The 2-iodo group precisely fits into a small sub-pocket shaped by the P2Y₁R's N terminus, and interacts with the main chain carbonyl of C42. This substituent has been shown to be critical for a high ligand-binding affinity to P2Y₁R. Derivatives containing 2-bromo, chloro and fluoro substitutions exhibit much lower affinities than MRS2500 (ref. 8). The N⁶-methyl group extends into another sub-pocket between helices VI and VII, forming hydrophobic interactions with A286[6.61] and N299[7.28]. The corresponding 6-amino analogue is 16-fold less potent, and any alkyl substitution
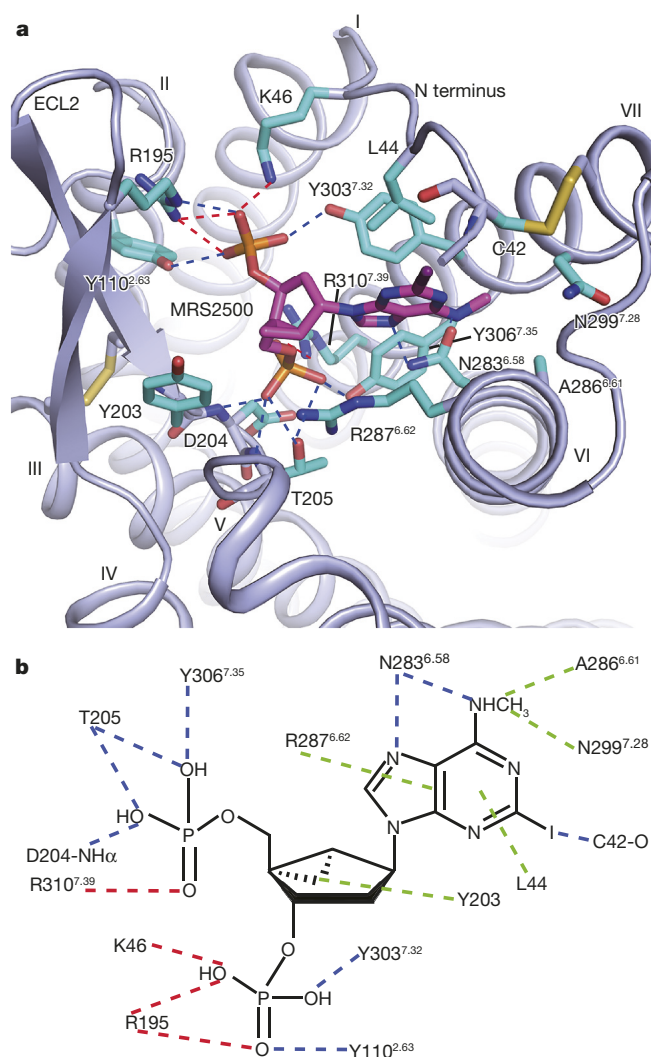
Figure 2 | P2Y₁R ligand-binding pocket for MRS2500. a, Key residues in P2Y₁R for MRS2500 binding. MRS2500 (magenta carbons) and P2Y₁R residues (cyan carbons) involved in ligand binding are shown in stick representation. The receptor is shown in blue cartoon representation. Other elements are coloured as follows: oxygen, red; nitrogen, dark blue; sulfur, yellow; phosphorus, orange; iodine, purple. Salt bridges are displayed as red dashed lines and hydrogen bonds as blue dashed lines. **b**, Schematic representation of interactions between P2Y₁R and MRS2500. Hydrophobic interactions are indicated as green dashed lines.

larger than ethyl abolishes binding to P2Y₁R[8]. The (N)-methanocarba ring makes a hydrophobic contact with the phenyl group of Y203 in ECL2. This is consistent with the observation that the Y203F mutation, but not Y203A, retains the ability to bind the nucleotide antagonist MRS2179 (ref. 19) (Extended Data Fig. 4). Both phosphate groups of nucleotide-like antagonists have been proven to be important for the high affinity interactions with P2Y₁R[20]. In the P2Y₁R–MRS2500 structure, each terminal oxygen of the two phosphates forms at least one contact with the receptor. The 3′-phosphate makes hydrogen bonds with Y110[2.63] and Y303[7.32] and is engaged in two salt-bridge interactions with K46 at the N terminus and R195 in ECL2. The 5′-phosphate forms a salt-bridge with R310[7.39] and makes hydrogen bonds with T205 in ECL2 and Y306[7.35].

Most of the above key residues of P2Y₁R have not been previously tested for their involvement in either agonist or antagonist binding. Our mutagenesis studies have further confirmed the critical roles of these residues in ligand binding of P2Y₁R, showing that the L44A, Y110[2.63]F, Y203A, T205A and N283[6.58]A mutants had greatly reduced binding

affinity for [³H]2-methylthio-adenosine 5′-diphosphate (2MeSADP), whereas the K46A, R195A, Y303[7.32]F mutants, surrounding the 3′-phosphate that is absent in 2MeSADP, selectively decreased the binding affinity of MRS2500 without affecting the binding of 2MeSADP and BPTU (Extended Data Table 2 and Extended Data Fig. 5). Removing the hydroxyl group coordinating the 5′-phosphate that is present in MRS2500 and 2MeSADP, Y306[7.35]F mutant displayed greatly reduced affinity of both nucleotides, consistent with a common binding site for both ligands. The requirement for T205 and N283[6.58] for 2MeSADP recognition is also consistent with a similar orientation of the both nucleotide ligands.

Previous mutagenesis of some residues in the helical bundle, such as H132[3.33], Y136[3.37], T222[5.43], F226[5.47] and K280[6.55] reduced P2Y$_1$R binding affinity of some agonist and antagonist ligands[20,21]. All of these residues are located much deeper than the MRS2500 binding site, but overlap well with the corresponding 2MeSADP binding site in the P2Y$_{12}$R structure. This observation raises the possibility of a second potential nucleotide binding site in the P2Y$_1$R receptor. Indeed, the P2Y$_1$R–MRS2500 structure has a deeper cavity that partially overlaps with the position of the adenine ring of 2MeSADP bound to P2Y$_{12}$R (Extended Data Fig. 6a). In the current P2Y$_1$R structure, however, the cavity is too small and the ECL2 extends deep into the ligand-binding pocket, blocking access to this potential binding site. Rearrangement of ECL2, as well as the helical bundle, would be required for a nucleotide to enter to this site. Nevertheless, it should be noted that the current data do not rule out the possibility that mutations of those residues may affect binding by changing the receptor conformation rather than by direct contact with a ligand.

The binding site of MRS2500 in P2Y$_1$R locates much closer to the extracellular surface than the small-molecule ligand-binding sites in the other known GPCR structures (Extended Data Fig. 6b). Comparing the antagonist binding sites in the three solved δ group class A GPCR structures, the binding sites of AZD1283 in P2Y$_{12}$R and vorapaxar in PAR1 are closer to helices IV and V than the MRS2500 binding site in P2Y$_1$R (Extended Data Fig. 6c). Although recognized by the same endogenous ligand ADP, P2Y$_1$R and P2Y$_{12}$R structures reveal very different features in binding their nucleotide-like ligands (Fig. 3). The ligand-binding sites for MRS2500 in P2Y$_1$R and 2MeSADP in P2Y$_{12}$R are spatially distinct, with only a minor overlap of phosphate binding regions near the residues at position 7.35. The adenine groups of the two ligands are in different orientations. In the P2Y$_1$R–MRS2500 structure, the adenine ring is adjacent to P2Y$_1$R's helices VI and VII, whereas the adenine group of 2MeSADP reaches deep into the binding pocket to form hydrophobic interactions with helices III and IV in the P2Y$_{12}$R structure.

A unique binding site for BPTU

The non-nucleotide ligand BPTU and other diarylurea P2Y$_1$R antagonists have been recently introduced as novel antiplatelet agents[10,22,23]. Surprisingly, the P2Y$_1$R–BPTU crystal structure reveals that instead of interacting within the seven-transmembrane helical bundle, BPTU binds to P2Y$_1$R on the lipidic interface of the transmembrane domain. The relatively shallow ligand-binding pocket, formed by aromatic and hydrophobic residues of helices I, II and III and ECL1, accommodates BPTU predominantly through hydrophobic interactions (Fig. 4 and Extended Data Fig. 3b). The only polar interactions are represented by
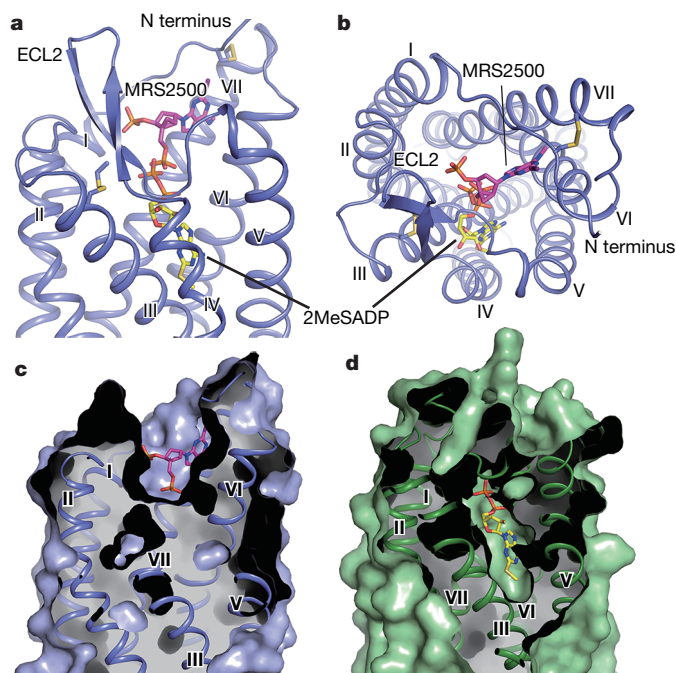
Figure 3 | Comparison of the ligand-binding modes between P2Y$_1$R–MRS2500 and P2Y$_{12}$R–2MeSADP. a, b, Side view (**a**) and top view (**b**) of the comparison of the ligand-binding sites between P2Y$_1$R–MRS2500 and P2Y$_{12}$R–2MeSADP. MRS2500 and 2MeSADP are shown in stick representation with magenta and yellow carbons, respectively. Only P2Y$_1$R represented as blue cartoon is shown. **c, d,** Comparison of the ligand-binding pockets between P2Y$_1$R–MRS2500 (**c**) and P2Y$_{12}$R-2MeSADP (**d**). The P2Y$_1$R and P2Y$_{12}$R structures are shown in cartoon and molecular surface representations, and coloured in blue and green, respectively.
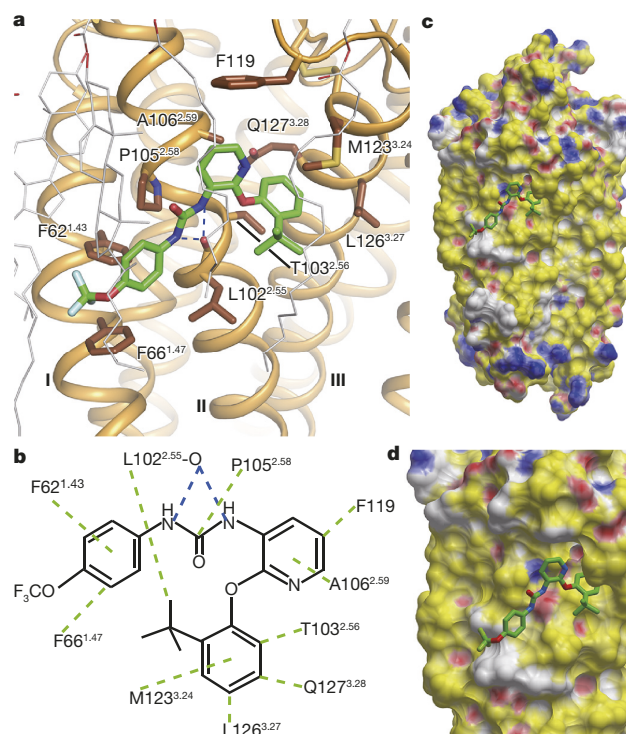
Figure 4 | P2Y$_1$R ligand-binding pocket for BPTU. a, Key residues in P2Y$_1$R for BPTU binding. BPTU (green carbons) and P2Y$_1$R residues (brown carbons) involved in ligand binding are shown in stick representation. The receptor is shown in orange cartoon representation. Lipid molecules in close contacts with BPTU are shown in white line representation. Hydrogen bonds are blue dashed lines. **b,** Schematic representation of interactions between P2Y$_1$R and BPTU. Hydrophobic interactions are green dashed lines. **c, d,** Side view (**c**) and zoom-in view (**d**) of the BPTU ligand-binding site in P2Y$_1$R. The BPTU binding site remains intact in the P2Y$_1$R-MRS2500 complex. The receptor is shown in molecular surface representation and coloured by binding property (yellow, hydrophobic surface; red, hydrogen bond acceptor potential; blue, hydrogen bond donor potential; white, neutral surface). The figure was prepared using ICM software (http://www.molsoft.com).

two hydrogen bonds between the nitrogen atoms of BPTU's urea group and the mainchain carbonyl of L102[2.55]. This carbonyl is available for bidentate coordination of this selective P2Y$_1$R antagonist because the residue P105[2.58] above precludes intrahelical hydrogen-bonding. Previous structure-activity relationship (SAR) studies demonstrated that replacing the urea linker of BPTU with other two to four atom linkers greatly reduced potency[24]. This could be explained by the importance of the two hydrogen-bond interactions for retaining P2Y$_1$R binding affinity of this chemical series. The pyridyl group forms hydrophobic interactions with A106[2.59] and F119, while its nitrogen atom is not involved in any interaction. However, if the pyridyl is substituted by a phenyl group, an extra hydrophobic contact with M123[3.24] may be introduced. This is supported by the fact that a corresponding phenyl derivative showed higher ligand-binding affinity and antiplatelet activity[22]. A106[2.59] is unique to P2Y$_1$ among P2YRs; other subtypes have larger side chains at this position, which could sterically hinder binding of BPTU's phenyl ring to this site, consistent with its P2Y$_1$R selectivity. Significantly, the sterically hindered A106[2.59]W/F/L mutants lost the ability to bind BPTU, while retaining recognition of nucleotide agonist 2MeSADP and antagonist MRS2500 (Extended Data Table 2 and Extended Data Fig. 5). The benzene ring within the phenoxy group of BPTU wedges into a cavity between helices II and III, interacting with T103[2.56], M123[3.24], L126[3.27] and Q127[3.28]. The hydrophobic nature of this sub-pocket is consistent with previous studies that have shown that the lipophilicity of this aryl group of the ligand is important for binding affinity and *in vitro* functional activity[22]. Similar to the A106[2.59]W mutant, the T103[2.56]W mutation abolished the binding affinity of BPTU to P2Y$_1$R, but did not affect the binding of 2MeSADP and MRS2500 (Extended Data Table 2). The *tert*-butyl substituent of the phenoxy ring forms a hydrophobic contact with L102[2.55] in the P2Y$_1$R-BPTU structure. SAR studies have shown that a lipophilic substitution of the phenoxy ring at either the *ortho* or *meta* position is preferred[10], consistent with the fact that a *para* substitution may cause spatial clashes with receptor's helix III based on the P2Y$_1$R–BPTU structure. At the other end of the ligand, the ureido phenyl ring forms two aromatic edge-to-face interactions with F62[1.43] and F66[1.47]. This aligns with the SAR requirement of aromatic monosubstitution of the urea N distal to the pyridine ring; replacing this aromatic ring with aliphatic substituents or inserting one or two methylene groups between a phenyl ring and the urea group decreased P2Y$_1$R binding affinity dramatically[10]. The above consistency between the P2Y$_1$R structure and the SAR studies has been validated also with docking simulations of several related, potent urea derivatives (Extended Data Fig. 7) that support the observed BPTU ligand-binding mode.

Previous efforts to reduce the lipophilicity of BPTU (HPLC log$P$ = 5.7) have shown that adding polar or basic amine groups to this ligand decreases binding affinity and antiplatelet activity; in fact, the binding affinity in this chemical series correlates directly with the HPLC log$P$ value[24]. These data agree with our structure where the ligand BPTU binds in a highly hydrophobic environment. The structural features of the ligand-binding site and the high lipophilicity of BPTU suggest that this ligand most likely enters the ligand-binding pocket via the lipid bilayer, rather than through the highly charged nucleotide approach route that leads to the P2Y$_1$R orthosteric site. The approach routes of hydrophobic GPCR ligands through the lipid bilayer have been proposed for several different receptors, such as rhodopsin, S1P1 and GPR40[25–27]. However, the ligands bound to these receptors at least partially occupy the conventional GPCR ligand-binding pocket, and to our knowledge, BPTU is the first structurally characterized selective and high affinity GPCR ligand that binds entirely outside of the helical bundle of GPCRs. The location of this ligand-binding site indicates that BPTU acts as an allosteric modulator of P2Y$_1$R. The allosteric regulation of P2Y$_1$R by BPTU was studied by investigating its ability to influence the dissociation of [^{3}H]2MeSADP from the receptor. The data indicate that BPTU can substantially accelerate the dissociation of 2MeSADP, while the

mutation A106[2.59]W abolishes the allosteric effect on the receptor by BPTU (Extended Data Fig. 8).

Insights into MRS2500 and BPTU inhibition mechanisms

Although MRS2500 and BPTU bind to distinct sites in P2Y$_1$R, these two ligands stabilize the receptor in similar inactive conformations. In comparing the active state structures of several different GPCRs with their inactive structures, a relatively stable bundle of helices I to IV, and a more mobile module consisting of helices V, VI and VII have been observed, demonstrating that the rearrangements of helices V, VI and VII play important roles in the process of GPCR activation[28–30]. In P2Y$_1$R, the antagonist MRS2500 potentially prevents such movements and stabilizes the receptor in an inactive state by interacting with helices VI and VII. In addition, it also bridges to the less mobile module through its 3′-phosphate, a requirement for antagonism in this chemical series. However, in the BPTU-bound P2Y$_1$R structure, the ligand makes no contacts with receptor's helices V, VI and VII, implying that BPTU inhibits receptor function in a different way. Besides the movements of helices V to VII, rotation and shift of helix III were also reported to be involved in the transformation from the inactive conformation to the active conformation of some GPCRs, although it is more subtle compared to the conformational changes of helices V to VII[17,28–30]. The P2Y$_1$R binding mode of BPTU suggests that this ligand most likely inhibits agonist-induced receptor activation by blocking relative movement of helices II and III, possibly a rotation of helix III. The above findings suggest that the movement of this less mobile module is equally critical for GPCR activation, in addition to the conformational changes of helices V to VII. Hindrance of the conformational plasticity of either domain may result in a severe loss of receptor function.

Conclusions

The P2Y$_1$R structures reveal atomic details of two completely distinct ligand-binding sites having chemically and structurally contrasting characteristics: a hydrophilic and charged site for the nucleotide-like antagonist MRS2500 with numerous anchor points on different receptor domains and a shallow site on the outer, lipid-exposed surface for the non-nucleotide ligand BPTU stabilized by a single carbonyl polar interaction. A comparison between P2Y$_1$R and P2Y$_{12}$R shows that these representative structures of two different P2YR subfamilies interact with their nucleotide ligands in disparate binding modes, which deepen our understanding of the diversity of signal recognition mechanisms in GPCRs. The external, hydrophobic binding site of BPTU suggests the entry of this allosteric antagonist occurs through the lipid membrane, and opens new opportunities to broaden the scope of future GPCR ligand discovery to target novel allosteric sites outside of the canonical ligand-binding pocket. The P2Y$_1$R structures also provide insights into the inhibition mechanisms of receptor activation by MRS2500 and BPTU through interactions with different domains of the receptor.

1. Abbracchio, M. P. *et al.* International Union of Pharmacology LVIII: update on the P2Y G protein-coupled nucleotide receptors: from molecular mechanisms and pathophysiology to therapy. *Pharmacol. Rev.* **58**, 281–341 (2006).
2. Gachet, C. P2 receptors, platelet function and pharmacological implications. *Thromb. Haemost.* **99**, 466–472 (2008).
3. Jacobson, K. A., Deflorian, F., Mishra, S. & Costanzi, S. Pharmacochemistry of the platelet purinergic receptors. *Purinergic Signal.* **7**, 305–324 (2011).
4. Jin, J. & Kunapuli, S. P. Coactivation of two different G protein-coupled receptors is essential for ADP-induced platelet aggregation. *Proc. Natl Acad. Sci. USA* **95**, 8070–8074 (1998).
5. Zerr, M. *et al.* Major contribution of the P2Y$_1$ receptor in purinergic regulation of TNFalpha-induced vascular inflammation. *Circulation* **123**, 2404–2413 (2011).

6. Fam, S. R., Gallagher, C. J. & Salter, M. W. P2Y$_1$ purinoceptor-mediated Ca^{2+} signaling and Ca^{2+} wave propagation in dorsal spinal cord astrocytes. *J. Neurosci.* **20,** 2800–2808 (2000).

7. Neary, J. T., Kang, Y., Willoughby, K. A. & Ellis, E. F. Activation of extracellular signal-regulated kinase by stretch-induced injury in astrocytes involves extracellular ATP and P2 purinergic receptors. *J. Neurosci.* **23,** 2348–2356 (2003).

8. Kim, H. S. *et al.* 2-Substitution of adenine nucleotide analogues containing a bicyclo[3.1.0]hexane ring system locked in a northern conformation: enhanced potency as P2Y$_1$ receptor antagonists. *J. Med. Chem.* **46,** 4974–4987 (2003).

9. Hechler, B. *et al.* MRS2500 [2-iodo-N6-methyl-(N)-methanocarba-2′-deoxyadenosine-3′,5′-bisphosphate], a potent, selective, and stable antagonist of the platelet P2Y$_1$ receptor with strong antithrombotic activity in mice. *J. Pharmacol. Exp. Ther.* **316,** 556–563 (2006).

10. Chao, H. *et al.* Discovery of 2-(phenoxypyridine)-3-phenylureas as small molecule P2Y$_1$ antagonists. *J. Med. Chem.* **56,** 1704–1714 (2013).

11. Zhang, C. *et al.* High-resolution crystal structure of human protease-activated receptor 1. *Nature* **492,** 387–392 (2012).

12. White, J. F. *et al.* Structure of the agonist-bound neurotensin receptor. *Nature* **490,** 508–513 (2012).

13. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330,** 1066–1071 (2010).

14. Tan, Q. *et al.* Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **341,** 1387–1390 (2013).

15. Wu, H. *et al.* Structure of the human kappa-opioid receptor in complex with JDTic. *Nature* **485,** 327–332 (2012).

16. Ballesteros, J. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).

17. Zhang, J. *et al.* Agonist-bound structure of the human P2Y$_{12}$ receptor. *Nature* **509,** 119–122 (2014).

18. Zhang, K. *et al.* Structure of the human P2Y$_{12}$ receptor in complex with an antithrombotic drug. *Nature* **509,** 115–118 (2014).

19. Costanzi, S., Mamedova, L., Gao, Z. G. & Jacobson, K. A. Architecture of P2Y nucleotide receptors: structural comparison based on sequence analysis, mutagenesis, and homology modeling. *J. Med. Chem.* **47,** 5393–5404 (2004).

20. Moro, S. *et al.* Human P2Y$_1$ receptor: molecular modeling and site-directed mutagenesis as tools to identify agonist and antagonist recognition sites. *J. Med. Chem.* **41,** 1456–1466 (1998).

21. Guo, D., von Kügelgen, I., Moro, S., Kim, Y. C. & Jacobson, K. A. Evidence for the recognition of non-nucleotide antagonists within the transmembrane domains of the human P2Y$_1$ receptor. *Drug Dev. Res.* **57,** 173–181 (2002).

22. Qiao, J. X. *et al.* Conformationally constrained ortho-anilino diaryl ureas: discovery of 1-(2-(1′-neopentylspiro[indoline-3,4′-piperidine]-1-yl)phenyl)-3-(4-(trifluoromethoxy)phenyl)urea, a potent, selective, and bioavailable P2Y$_1$ antagonist. *J. Med. Chem.* **56,** 9275–9295 (2013).

23. Yang, W. *et al.* Discovery of 4-aryl-7-hydroxyindoline-based P2Y$_1$ antagonists as novel antiplatelet agents. *J. Med. Chem.* **57,** 6150–6164 (2014).

24. Wang, T. C. *et al.* Discovery of diarylurea P2Y$_1$ antagonists with improved aqueous solubility. *Bioorg. Med. Chem. Lett.* **23,** 3239–3243 (2013).

25. Hildebrand, P. W. *et al.* A ligand channel through the G protein coupled receptor opsin. *PLoS ONE* **4,** e4382 (2009).

26. Hanson, M. A. *et al.* Crystal structure of a lipid G protein-coupled receptor. *Science* **335,** 851–855 (2012).

27. Srivastava, A. *et al.* High-resolution structure of the human GPR40 receptor bound to allosteric agonist TAK-875. *Nature* **513,** 124–127 (2014).

28. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–187 (2008).

29. Rasmussen, S. G. *et al.* Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* **469,** 175–180 (2011).

30. Xu, F. *et al.* Structure of an agonist-bound human A$_{2A}$ adenosine receptor. *Science* **332,** 322–327 (2011).

Author Contributions D.Z. optimized the construct, developed the purification procedure and purified the P2Y$_1$R proteins for crystallization, performed crystallization trials and optimized crystallization conditions. Z.-G.G. designed, performed and analysed ligand binding and competition assays of wild-type and mutant P2Y$_1$R. K.Z. helped with construct and crystal optimization, and collected diffraction data. E.K. and J.W. helped with ligand synthesis of P2Y$_1$R. S.C. performed and analysed ligand-binding assays. S.P. performed and analysed docking assays. C.Y. and L.M. expressed the P2Y$_1$R proteins. W.Z. developed the initial expression and purification protocol for P2Y$_1$R. G.W.H. helped to analyse the structures. H.L. oversaw ligand synthesis of P2Y$_1$R. V.C. and V.K. helped to analyse the structures and assisted with manuscript preparation. H.J. and R.C.S. oversaw structure analysis/interpretation of P2Y$_1$R. K.A.J. oversaw, designed and analysed ligand-binding assays, oversaw ligand synthesis, and assisted with manuscript preparation. Q.Z. and B.W. initiated the project, planned and analysed experiments, solved the structures, supervised the research and wrote the manuscript.

METHODS

Cloning and expression of engineered P2Y$_1$R proteins. The wild-type (WT) human P2Y$_1$R DNA was synthesized by Genewiz and then cloned into a modified pFastBac1 vector (Invitrogen) containing an expression cassette with an HA signal sequence followed by a Flag tag at the N terminus and a PreScission protease site followed by a 10× His tag at the C terminus. An engineered construct (construct 1) was generated by overlap extension PCR to insert M1–E54 of rubredoxin[31] between K247 and P253 in the intracellular loop 3 (ICL3) of P2Y$_1$R. The P2Y$_1$R gene was further modified by introducing the D320$^{7.49}$N mutation to improve protein yield and stability. Another P2Y$_1$R construct (construct 2) was made by adding A23–L128 of a thermostabilized BRIL (PDB ID: 1M6T) before the residue A8 of the receptor sequence in construct 1. High-titre recombinant baculovirus (>10^8 viral particles per ml) was obtained using the Bac-to-Bac Baculovirus Expression System (Invitrogen). *Spodoptera frugiperda* (*Sf*9) cells at cell density of 2×10^6 to 3×10^6 cells per ml were infected with virus at MOI (multiplicity of infection) of 5. Cells were collected by centrifugation at 48 h post-infection and stored at −80 °C until use.

Purification of *Sf*9-expressed P2Y$_1$R proteins for crystallization. Insect cell membranes were disrupted by thawing frozen cell pellets in a hypotonic buffer containing 10 mM HEPES, pH 7.5, 10 mM MgCl$_2$, 20 mM KCl and EDTA-free complete protease inhibitor cocktail (Roche). Cell membranes were disrupted by repeated dounce homogenization. Extensive washing of the membranes was performed by centrifugation in the same hypotonic buffer (one more time), followed by a high osmotic buffer containing 1 M NaCl, 10 mM HEPES, pH 7.5, 10 mM MgCl$_2$, 20 mM KCl (three times), thereby removing soluble and membrane associated proteins from the suspension of membranes, and then the hypotonic buffer (one more time) to remove the high concentration of NaCl. Purified membranes were resuspended in 10 mM HEPES, pH 7.5, 30% (v/v) glycerol, 10 mM MgCl$_2$, 20 mM KCl and EDTA-free complete protease inhibitor cocktail, flash-frozen with liquid nitrogen, and stored at −80 °C until further use.

Prior to solubilization, the purified membranes of construct 1-expressed materials were thawed on ice in the presence of 1 mM ATP (Sigma), 2 mg ml^{-1} iodoacetamide (Sigma) and EDTA-free protease inhibitor cocktail (Roche). After incubating at 4 °C for 1 h, the membranes were then solubilized in 50 mM HEPES, pH 7.5, 300 mM NaCl, 0.5% (w/v) n-dodecyl-β-D-maltopyranoside (DDM, Anatrace), 0.1% (w/v) cholesterol hemisuccinate (CHS) (Sigma), and 500 μM ATP for 3 h at 4 °C. The supernatant was isolated by centrifugation at 160,000g for 30 min, supplemented with 30 mM imidazole, pH 7.5, and incubated with TALON IMAC resin (Clontech) overnight at 4 °C. The resin was washed with ten column volumes of 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 40 mM imidazole, 0.05% (w/v) DDM, 0.01% (w/v) CHS and 1 mM ATP, followed by ten column volumes of 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10 mM MgCl$_2$, 5 mM ATP, and fifteen column volumes of 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 0.05% (w/v) CHS and 50 μM MRS2500. The protein was eluted by 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 300 mM imidazole, 0.05% (w/v) DDM, 0.01% (w/v) CHS and 100 μM MRS2500 in five column volumes. PD MiniTrap G-25 column (GE healthcare) was used to remove imidazole and increase the compound concentration to 1 mM. The protein was then treated overnight with His-tagged PreScission protease (home-made) and His-tagged PNGase F (home-made) to remove the C-terminal His-tag and de-glycosylate the receptor. PreScission protease, PNGaseF and the cleaved His tag were removed by Ni-NTA superflow resin (Qiagen) incubation at 4 °C for 1 h. The His-tag cleaved protein was collected in the Ni-NTA column flow through, and then concentrated to 40–50 mg ml^{-1} with a 100 kDa molecular weight cut-off Vivaspin concentrator (Sartorius Stedim Biotech). Receptor purity, monodispersity and concentration were estimated using SDS–PAGE and analytical size-exclusion chromatography (aSEC).

The P2Y$_1$R–BPTU complex protein was purified following a protocol similar to the above procedure. The membranes of construct 2-expressed materials were incubated in 50 μM BPTU, 2 mg ml^{-1} iodoacetamide (Sigma), and EDTA-free protease inhibitor cocktail (Roche) at 4 °C for 1 h, and then solubilized in 50 mM HEPES, pH 7.5, 300 mM NaCl, 0.5% (w/v) DDM, 0.1% (w/v) CHS and 25 μM BPTU for 3 h at 4 °C. After overnight binding to the TALON IMAC resin, the resin was washed with thirty column volumes of 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 40 mM imidazole, 0.05% (w/v) DDM, 0.01% (w/v) CHS and 25 μM BPTU, and then eluted by five column volumes of 25 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 300 mM imidazole, 0.05% (w/v) DDM, 0.01% (w/v) CHS and 50 μM BPTU. Imidazole was removed using the PD MiniTrap G-25 column. The protein was further purified by the treatment with His-tagged PreScission protease and PNGase F.

Lipidic cubic phase crystallization of P2Y$_1$R–MRS2500 and P2Y$_1$R–BPTU complexes. Purified protein samples of P2Y$_1$R were reconstituted into lipidic cubic phase (LCP) by mixing with molten lipid in a mechanical syringe mixer[32]. The protein solution was mixed with monoolein/cholesterol (10:1 by mass) lipids at weight

ratio of 1:1.5 (protein: lipid). After formation of a transparent lipidic cubic phase, the mixture was dispensed onto 96-well glass sandwich plates (Shanghai FAstal BioTech) in 40–50 nl drops and overlaid with 800 nl precipitant solution using a Mosquito LCP robot (TTP Labtech). Protein reconstitution in LCP and crystallization trials were performed at room temperature (19–22 °C). Plates were incubated and imaged at 20 °C using an automated incubator/imager (RockImager, Formulatrix). The crystals of the P2Y$_1$R–MRS2500 complex grew to their full size (70–150 μm) within two weeks in 20–30% PEG400 (v/v), 50–100 mM sodium citrate, 50 μM MRS2500, and 0.1 M HEPES, pH 7.0 or 0.1 M Tris-HCl, pH 8.0. The P2Y$_1$R–BPTU complex was crystallized in 100–300 mM ammonium phosphate dibasic, 0–10% PEG2000 MME, 50 μM BPTU, and 0.1 M sodium citrate, pH 6.5, and the crystals reached their maximum size (100–130 μm) within two weeks. The P2Y$_1$R crystals were collected directly from LCP using 50–100 μm micromounts (M2-L19-50/100, MiTeGen) and flash frozen in liquid nitrogen.

Data collection and structure determination. X-ray diffraction data were collected at the SPring-8 beam line 41XU, Hyogo, Japan, using a Pilatus3 6M detector (X-ray wavelength 1.0000 Å). The crystals were exposed with a 10 μm mini-beam for 0.5 s and 0.5° oscillation per frame. XDS[33] was used for integrating and scaling data from 36 best-diffracting crystals of the P2Y$_1$R–MRS2500 complex and 12 crystals of the P2Y$_1$R–BPTU complex. Initial phase information of the P2Y$_1$R–MRS2500 complex was obtained by molecular replacement (MR) with Phaser[34] using the receptor portion of PAR1 (PDB ID: 3VW7), converted to polyalanines, and rubredoxin structure (PDB ID: 1IRO) as search models. The correct MR solution contained two P2Y$_1$R–rubredoxin molecules packed antiparallel in the asymmetric unit. Refinement was performed with REFMAC5[35] and BUSTER[36] followed by manual examination and rebuilding of the refined coordinates in the program COOT[37] using both $|2F_o| - |F_c|$ and $|F_o| - |F_c|$ maps. The final model includes 296 residues (38–247 and 253–338) of the 345 residues of P2Y$_1$R and residues 1 to 54 of rubredoxin. The remaining N- and C-terminal residues are disordered and were not refined. The P2Y$_1$R–BPTU complex structure was solved using P2Y$_1$R in the P2Y$_1$R–MRS2500 complex and rubredoxin as starting models and refined under the same procedure. The final model of the P2Y$_1$R–BPTU complex contains 291 residues (38–156, 158–247 and 253–334) of P2Y$_1$R and the 54 residues of rubredoxin. Without clear electronic density, the N-terminal fused BRIL was not traced. The crystal packing shows that there is no room to fit BRIL, indicating that it most likely degraded.

Ligand-binding assays. Materials: 2MeSADP and MRS2500 were from Tocris (Minneapolis, USA). [^{3}H]2MeSADP (3.5 Ci per mmol) was purchased from Moravek, Brea, USA. BPTU was synthesized as reported using a 2-aryloxy-3-isothiocyana-topyridine as intermediate[10].

Membrane preparations from Sf9 cells and COS-7 cells expressing WT and mutant human P2Y$_1$Rs were used for all the ligand-binding assays. Protein concentrations were measured using Bio-Rad protein assay reagents. For saturation experiments, 50 μl [^{3}H]2MeSADP (from 5 to 200 nM) was incubated with 100 μl WT and mutant P2Y$_1$R membrane preparations (5 μg per tube) in a total assay volume of 200 μl Tris-HCl buffer containing 10 mM MgCl$_2$. MRS2500 or 2MeSADP (10 μM) was used to determine the non-specific binding. For displacement experiments using the membrane preparations from Sf9 cells, increasing concentrations of MRS2500 or BPTU were incubated with WT or mutant P2Y$_1$R membrane preparations (5–10 μg) and 25 nM [^{3}H]2MeSADP at 25 °C for 30 min. Using the membrane preparations from COS-7 cells, increasing concentrations of MRS2500 or BPTU were incubated with WT or mutant P2Y$_1$R membrane preparations (20 μg) and 2 nM [^{3}H]2MeSADP at 25 °C for 30 min.

For dissociation experiments, WT or mutant P2Y$_1$R membrane preparations from Sf9 cells (5 μg) were first pre-equilibrated with 25 nM [^{3}H]2MeSADP at 4 °C for 30 min. Using the membrane preparations from COS-7 cells, membranes containing 20 μg of protein were pre-equilibrated with 2 nM [^{3}H]2MeSADP at 4 °C for 30 min. Then the dissociation was initiated at 4 °C by mixing with 10 μM MRS2500 in the absence or presence of 10 μM BPTU (note that the dissociation at 25 °C was too fast to be measured). The reaction was terminated by harvesting with a 24-channnel Brandel cell harvester (Brandel, Gaithersburg, USA) and followed by washing twice with 5 ml cold Tris-HCl buffer containing 10 mM MgCl$_2$. Radioactivity was measured using a scintillation counter (Tri-Carb 2810TR). Data were analysed using Prism 6 (GraphPad, San Diego, USA).

Docking simulations of BPTU derivatives. The P2Y$_1$R-BPTU structure was prepared using the Protein Preparation Wizard tool implemented in the Schrödinger suite, adding all the hydrogen atoms and the missing side chains of residues whose backbone coordinates were observed in the structure. The orientation of polar hydrogens was optimized, the protein protonation states were adjusted and the overall structure was minimized with harmonic restraints on the heavy atoms, to remove strain. Then, all the hetero groups and water molecules were deleted.

The SiteMap tool of the Schrödinger suite was used to identify potential binding sites in the structure. In addition to the canonical orthosteric binding site within

the transmembrane bundle, a shallow pocket was identified on the external receptor interface with the lipid bilayer in correspondence of the BPTU crystallographic pose and was selected as the docking site. Molecular docking of several BPTU derivatives at the P2Y$_1$R structure was performed by means of the Glide package from the Schrödinger suite. In particular, a Glide Grid was centred on the centroid of residues located within 5 Å from the previously identified cavity. The Glide Grid was built using an inner box (ligand diameter midpoint box) of 14 Å × 14 Å × 14 Å and an outer box (within which all the ligand atoms must be contained) that extended 20 Å in each direction from the inner one. Docking of ligands was performed in the rigid binding site using the SP (standard precision) procedure. The top scoring docking conformations accurately reproduced the binding mode observed for BPTU in the crystal and were in agreement with previous SAR findings for this class of compounds.

Sample size. No statistical methods were used to predetermine sample size.

31. Chun, E. *et al.* Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* **20,** 967–976 (2012).
32. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
33. Kabsch, W. Xds. *Acta Crystallogr. D* **66,** 125–132 (2010).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
35. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
36. Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. D* **68,** 368–380 (2012).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).

LETTER

Nuclear ashes and outflow in the eruptive star Nova Vul 1670

Tomasz Kamiński[1,2], Karl M. Menten[2], Romuald Tylenda[3], Marcin Hajduk[3], Nimesh A. Patel[4] & Alexander Kraus[2]

CK Vulpeculae was observed in outburst in 1670–1672 (ref. 1), but no counterpart was seen until 1982, when a bipolar nebula was found at its location[1–3]. Historically, CK Vul has been considered to be a nova (Nova Vul 1670), but its similarity to 'red transients', which are more luminous than classical novae and thought to be the results of stellar collisions[4], has re-opened the question of CK Vul's status[5,6]. Red transients cool to resemble late M-type stars, surrounded by circumstellar material rich in molecules and dust[7–9]. No stellar source has been seen in CK Vul, though a radio continuum source was identified at the expansion centre of the nebula[3]. Here we report that CK Vul is surrounded by chemically rich molecular gas in the form of an outflow, as well as dust. The gas has peculiar isotopic ratios, revealing that CK Vul's composition was strongly enhanced by the nuclear ashes of hydrogen burning. The chemical composition cannot be reconciled with a nova or indeed any other known explosion. In addition, the mass of the surrounding gas is too large for a nova, though the conversion from observations of CO to a total mass is uncertain. We conclude that CK Vul is best explained as the remnant of a merger of two stars.

Using the submillimetre-wave Atacama Pathfinder Experiment (APEX) telescope, located in the Chilean Andes, we discovered bright and chemically complex molecular gas in emission, which has not been observed before in CK Vul. A spectral-line survey in the 217–910 GHz range revealed emission from a plethora of molecules (Fig. 1a–c). Additionally, using the Effelsberg radio telescope, we observed inversion lines of NH_3 (Fig. 1d). The detected transitions are listed in Extended Data Table 1. Our excitation analysis indicates that the molecular gas is cool, with rotational temperatures of 8–22 K, but some amount of gas at higher excitation is also evident.

The molecular inventory implies that the abundance of nitrogen is greatly enhanced in CK Vul. The paucity of oxides, namely the lack of SO, SO_2, and maser emission of H_2O, and OH (typically omnipresent in oxygen-rich environments) implies that the circumstellar material is not dominated by oxygen. We do observe relatively strong lines from some species containing oxygen—SiO, CO, HCO^+, and H_2CO—but those molecules are also observed in envelopes of carbon-rich stars. The gas does not appear to be carbon-rich because many species typical for such environments (for example, SiC, SiC_2, and HC_3N), although covered by our spectra, are not observed. Again, all the carbon-bearing molecules present in CK Vul, that is, CO, CS, H_2CO, and HCO^+, have been observed in other chemical types of circumstellar envelopes[10]. What is unusual in CK Vul is a rich variety of nitrogen-bearing species. Of all the nitrogen-bearing species predicted at thermal equilibrium to be abundant in gas greatly enhanced in nitrogen[11], only NO and N_2
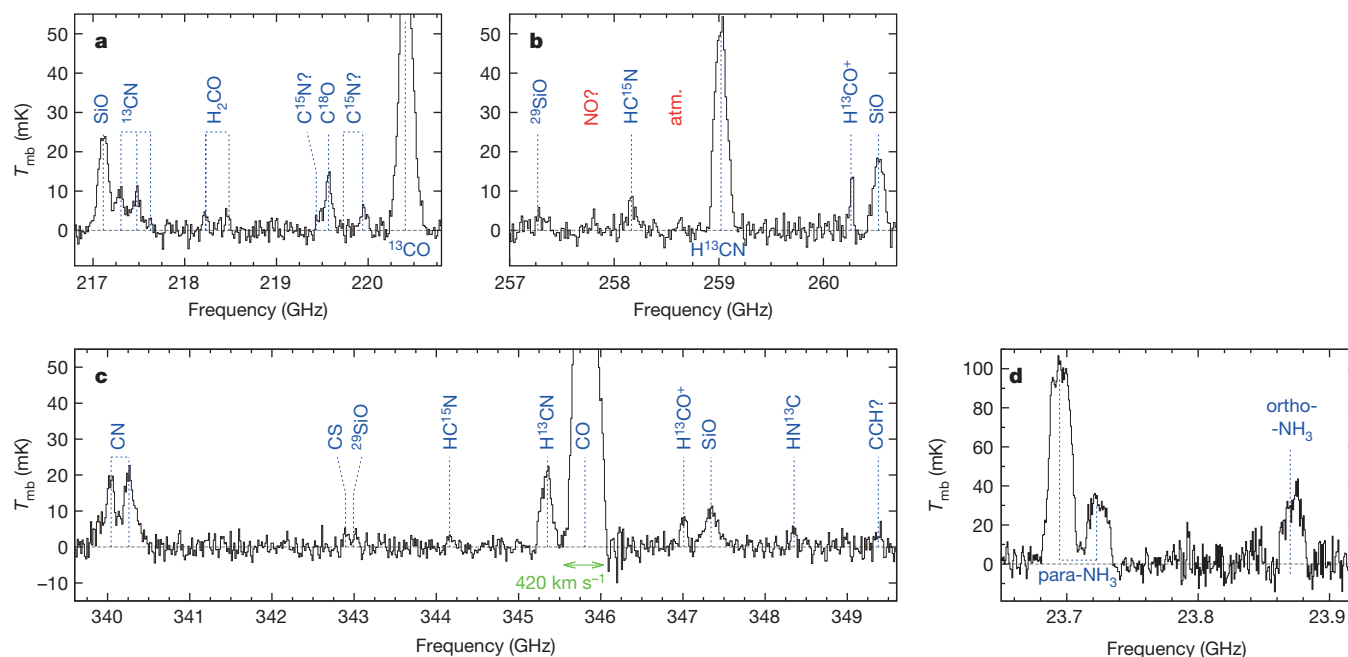
Figure 1 | Spectra of CK Vul with identifications of the main features.
a–c, Example APEX spectra containing some of the observed emission lines. Lines of CO, CS, SiO, CN, HCN, HNC, HCO^+, N_2H^+, H_2CO, and their isotopologues (that is, isotopic variations) were observed. d, The Effelsberg spectrum of the inversion lines of ammonia. The lines show very broad profiles with full widths of up to ~420 km s^{-1}. Some telluric residuals are present in the spectra, the strongest of which is marked as 'atm.'

[1]European Southern Observatory, Alonso de Córdova 3107, Vitacura, Santiago, Chile. [2]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany. [3]Department for Astrophysics, N. Copernicus Astronomical Center, Rabiańska 8, 87-100, Toruń, Poland. [4]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.
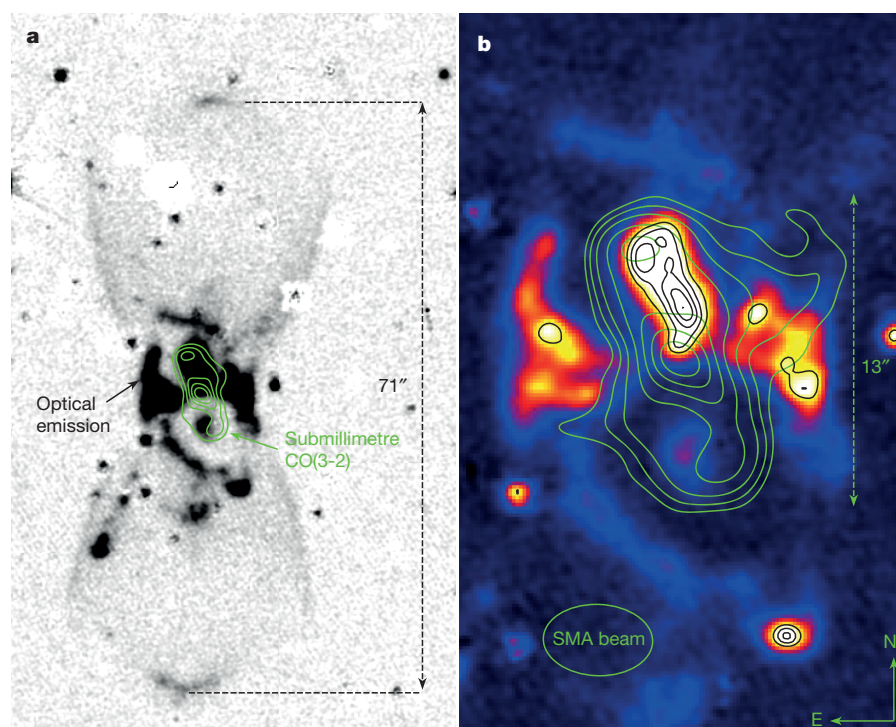
Figure 2 | The ionized nebula and the newly discovered molecular emission in CK Vul. a, The image shows the Hα + [N II] nebula created in the seventeenth-century explosion. Bright stars were removed from this optical image[3]. Green contours show the emission in the ^{12}CO $J = 3$–2 transition observed at submillimetre wavelengths (at 29%, 43%, 57%, 72% and 86% of the maximum emission). **b**, The central part of the nebula is shown in colour scale with yellow showing the brightest parts, and blue, faint emission. The structure of the bright optical jet is shown with black contours. Two extra green contours are drawn for CO emission, at 12% and 20% of the peak intensity. Dashed lines show the scale.

remain undetected in CK Vul. The N_2 molecule has no allowed rotational transitions, while transitions of NO, although covered by APEX spectra, might have been undetected owing to the low levels of oxygen and the small dipole moment of the NO molecule. All the nitrogen-bearing species observed in CK Vul are also present in the envelopes of the yellow supergiant IRC+10420 and the luminous blue variable η Carinae, both of which were recently proposed to be prototypes of nitrogen-rich objects[11,12]. This makes CK Vul only the third known such case. An overabundance of nitrogen relative to oxygen had been suggested for CK Vul, based on observations of the optical atomic lines[1], but the result was questionable owing to uncertain assumptions.

Some of the transitions covered by APEX were later observed at higher angular resolution with the Submillimetre Array (SMA), an interferometer located in Hawaii. The maps reveal that the emission arises from a bipolar structure about 15″ in size. This molecular region is much smaller than the long-known ionized nebula (extending over about 71″; Fig. 2a). The spatio-kinematical structure of the lobes is complex and suggests the presence of two partially overlapping hourglass-shaped shells observed at very low inclination angles. The lobes are apparent only in some of the observed transitions; most of the molecular emission arises in the central source, which is only partially resolved at our best resolution of approximately 2″. The northern molecular lobe coincides very closely with the brightest clump of the optical nebula (Fig. 2b), suggesting that the molecular gas coexists with the plasma. The observed misalignment of the long axis of the molecular region with respect to the axis of the large-scale optical nebula might be caused by precession.

In addition to molecular lines, continuum emission was observed with the SMA, revealing thermal emission of dust arising from the position where radio continuum was found in earlier observations[3]. The millimetre-wave source is dominated by a structure 3.7″ × 1.0″ in size and seen at the position angle of 33.4°, but also has components extending a few arcseconds along the northern and southern molecular lobes. The continuum indicates the presence of a flattened dusty envelope, perhaps a torus, and a pair of collimated jets. Our analysis of all available continuum measurements, ranging from micrometre to centimetre wavelengths, indicates that the emission is dominated by dust at a temperature of around 15 K but warmer dust up to 50 K must also be present.

From our rough estimate of the carbon monoxide (CO) column density, 4×10^{17} cm^{-2}, we calculate the total mass of the gas to be about one solar mass. Here we assumed that the CO abundance with respect to hydrogen is of the order of 10^{-4}, as found in many interstellar/circumstellar environments of various types. The possible line-saturation effects would make our estimate a lower limit on the total mass. The peculiar elemental composition of CK Vul indicates, however, that the CO abundance may deviate from the classical value. If the overabundance of nitrogen is owing to its production at the cost of carbon and oxygen, the actual CO abundance with respect to hydrogen could be lower than 10^{-4} and then our value underestimates the total mass of the gas. In the case of strong enrichment of helium at the expense of hydrogen, our estimate should be reasonably close to the actual total mass, because the correction for the presence of helium would compensate for the deficiency of hydrogen. Also, the mass should be enlarged by the contribution of the material seen in the optical nebula, a number which remains unknown. Our mass estimate, although uncertain, is much higher than the mass that a classical nova explosion is able to accumulate during its lifetime[13].

The presence of the strong submillimetre-wave molecular emission itself makes CK Vul an extraordinary eruptive variable star. Classical novae do not show such emission, as we recently confirmed by observing 17 Galactic-disk novae with APEX. Galactic red transients, which have rich molecular spectra at optical and near-infrared wavelengths, have also not been detected in submillimetre-wave thermally excited emission lines[8,14].

In fact, the central object of CK Vul may be hostile to molecules, as suggested by the presence of the ionic species HCO$^+$ and N_2H^+. Their formation channels in the absence of water require a high abundance of H_3^+, which can be formed from H_2 exposed to an ultraviolet radiation field[15] or by shocks. The high outflow velocity of $\sim$210 km s^{-1} observed in CK Vul, the presence of jets, and emission of atomic ions[3,16] make shocks a more favourable ionization mechanism.

There is a striking resemblance of the newly revealed observational characteristics of CK Vul to those of a short evolutionary stage of low- to intermediate-mass stars known as preplanetary nebulae, especially to OH231.8+4.2 (the Calabash Nebula), which has an extended pair of lobes seen in optical atomic lines and a pair of molecular jets emanating

Table 1 | Isotopic ratios of the molecular gas of CK Vul

Isotopologues	Column-density ratio of the isotopologue pairs
$^{12}C^{16}O/^{13}C^{16}O$	6 ± 2
$^{12}C^{16}O/^{12}C^{18}O$	23 ± 15
$^{12}C^{16}O/^{12}C^{17}O$	$\gg 225$
$H^{12}C^{14}N/H^{13}C^{14}N$	3 ± 1
$H^{12}C^{14}N/H^{12}C^{15}N$	26 ± 9
$^{12}C^{14}N/^{13}C^{14}N$	~ 2
$^{12}C^{14}N/^{12}C^{15}N$	$\sim 4*$
$H^{12}CO^+/H^{13}CO^+$	2 ± 1
$^{28}SiO/^{29}SiO$	4 ± 4

* Based on uncertain identification.

from a dusty flattened structure[17]. At least some of the known preplanetary nebulae must have been formed in a short and energetic event[18,19] and it has recently been proposed that the type of explosions we have witnessed in red transients may be actually responsible for the formation of the circumstellar material of preplanetary nebulae[18,20]. Our observations of CK Vul would then provide strong support for such a link.

However, our analysis leads to the conclusion that the remnant of Nova Vul 1670 must be of a different nature to that of preplanetary nebulae. First of all, its spectral energy distribution (Extended Data Fig. 1) implies a luminosity of around 0.9 solar luminosities, while preplanetary nebulae reach luminosities of the order of 10,000 solar luminosities. Moreover, the chemical composition of CK Vul, especially the nitrogen enrichment, would be very unusual for a preplanetary nebula. Also anomalous is the presence of lithium in the outflow of CK Vul, as evidenced by two variable field stars whose spectra show absorption lines of lithium[16].

The strongest argument for CK Vul being a unique transient comes from our analysis of its isotopic abundances. The column density ratios of different isotopologues listed in Table 1 probably represent the true isotopic ratios of the different elements (but may be somewhat influenced by photo-chemical fractionation and opacity effects). The isotopic ratios of the CNO elements compared to solar values[21] (in parentheses)—that is, $^{12}C/^{13}C = 2$–6 (solar 89), $^{14}N/^{15}N \approx 26$ (solar 272), $^{16}O/^{18}O \approx 23$ (solar 499), and $^{16}O/^{17}O > 225$ (solar 2,682)—reveal a very peculiar isotopic pattern that undoubtedly indicates nuclear processing of the circumstellar gas. The pattern could not be produced by an asymptotic-giant-branch star or a post-asymptotic-giant-branch /preplanetary nebula object because these are characterized by much higher ratios of $^{16}O/^{18}O$ and $^{14}N/^{15}N$ (ref. 22); in fact, isotopologues containing ^{15}N are never observed in spectra of those evolved stars. The isotopic ratios obtained cannot be reconciled with the current understanding of thermonuclear runaway nucleosynthesis, mainly because nova ashes have a much lower $^{16}O/^{17}O$ ratio[23].

It is most tempting to consider that CK Vul underwent its seventeenth-century cataclysm owing to a merger of stars, given that such events have now been proved to explain the explosions of red transients[4]. The explosion could have been violent enough to penetrate and eject inner parts of the merging stars, exposing material that was active in nuclear burning. Interestingly, the general elemental abundances revealed by the molecular spectra here are well reproduced by abundances expected for non-explosive hydrogen burning in the CNO cycles[24]. Not all of the observed isotopic signatures fit those models, though, with the observed ratios of $^{15}N/^{14}N$ and $^{16}O/^{18}O$ being too high. However, a merger remnant could be a complex mixture of processed and unprocessed gas and no quantitative predictions exist for the chemical composition of such an exotic star and its circumstellar environment.

Interestingly, the $^{12,13}C$ and $^{14,15}N$ isotopic ratios of CK Vul are close to those of presolar grains known as 'nova grains'[25] but of unclear origin[26,27]. Although the agreement in abundances is conspicuous, those stardust grains originate from carbon-rich environments, which makes their link to the CK Vul phenomenon elusive.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Shara, M. M., Moffat, A. F. J. & Webbink, R. F. Unraveling the oldest and faintest recovered nova—CK Vulpeculae (1670). *Astrophys. J.* **294,** 271–285 (1985).
2. Shara, M. M. & Moffat, A. F. J. The recovery of CK Vulpeculae Nova 1670—the oldest 'old nova'. *Astrophys. J.* **258,** L41–L44 (1982).
3. Hajduk, M. *et al.* The enigma of the oldest 'nova': the central star and nebula of CK Vul. *Mon. Not. R. Astron. Soc.* **378,** 1298–1308 (2007).
4. Tylenda, R. *et al.* V1309 Scorpii: merger of a contact binary. *Astron. Astrophys.* **528,** A114 (2011).
5. Kato, T. CK Vul as a candidate eruptive stellar merging event. *Astron. Astrophys.* **399,** 695–697 (2003).
6. Tylenda, R. *et al.* OGLE-2002-BLG-360: from a gravitational microlensing candidate to an overlooked red transient. *Astron. Astrophys.* **555,** A16 (2013).
7. Kamiński, T., Schmidt, M., Tylenda, R., Konacki, M. & Gromadzki, M. Keck/HIRES spectroscopy of V838 Monocerotis in October 2005. *Astrophys. J. Suppl. Ser.* **182,** 33–50 (2009).
8. Kamiński, T., Schmidt, M. & Tylenda, R. V4332 Sagittarii: a circumstellar disc obscuring the main object. *Astron. Astrophys.* **522,** A75 (2010).
9. Nicholls, C. P. *et al.* The dusty aftermath of the V1309 Sco binary merger. *Mon. Not. R. Astron. Soc.* **431,** L33–L37 (2013).
10. Ziurys, L. M., Tenenbaum, E. D., Pulliam, R. L., Woolf, N. J. & Milam, S. N. Carbon chemistry in the envelope of VY Canis Majoris: implications for oxygen-rich evolved stars. *Astrophys. J.* **695,** 1604–1613 (2009).
11. Quintana-Lacaci, G. *et al.* Detection of circumstellar nitric oxide. Enhanced nitrogen abundance in IRC +10420. *Astron. Astrophys.* **560,** L2 (2013).
12. Loinard, L., Menten, K. M., Güsten, R., Zapata, L. A. & Rodríguez, L. F. Molecules in η Carinae. *Astrophys. J.* **749,** L4 (2012).
13. Romano, D. & Matteucci, F. Nova nucleosynthesis and Galactic evolution of the CNO isotopes. *Mon. Not. R. Astron. Soc.* **342,** 185–198 (2003).
14. Kamiński, T. Extended CO emission in the field of the light echo of V838 Monocerotis. *Astron. Astrophys.* **482,** 803–808 (2008).
15. Mamon, G. A., Glassgold, A. E. & Omont, A. Photochemistry and molecular ions in oxygen-rich circumstellar envelopes. *Astrophys. J.* **323,** 306–315 (1987).
16. Hajduk, M., van Hoof, P. A. M. & Zijlstra, A. A. CK Vul: evolving nebula and three curious background stars. *Mon. Not. R. Astron. Soc.* **432,** 167–175 (2013).
17. Bujarrabal, V., Alcolea, J., Sánchez Contreras, C. & Sahai, R. HST observations of the protoplanetary nebula OH 231.8+4.2: the structure of the jets and shocks. *Astron. Astrophys.* **389,** 271–285 (2002).
18. Soker, N. & Kashi, A. Formation of bipolar planetary nebulae by intermediate-luminosity optical transients. *Astrophys. J.* **746,** 100 (2012).
19. Szyszka, C., Zijlstra, A. A. & Walsh, J. R. The expansion proper motions of the planetary nebula NGC 6302 from Hubble Space Telescope imaging. *Mon. Not. R. Astron. Soc.* **416,** 715–726 (2011).
20. Prieto, J. L., Sellgren, K., Thompson, T. A. & Kochanek, C. S. A. Spitzer/IRS spectrum of the 2008 luminous transient in NGC 300: connection to proto-planetary nebulae. *Astrophys. J.* **705,** 1425–1432 (2009).
21. Lodders, K. Solar system abundances and condensation temperatures of the elements. *Astrophys. J.* **591,** 1220–1247 (2003).
22. Kobayashi, C., Karakas, A. I. & Umeda, H. The evolution of isotope ratios in the Milky Way Galaxy. *Mon. Not. R. Astron. Soc.* **414,** 3231–3250 (2011).
23. Denissenkov, P. A. *et al.* MESA and NuGrid simulations of classical novae: CO and ONe nova nucleosynthesis. *Mon. Not. R. Astron. Soc.* **442,** 2058–2074 (2014).
24. Arnould, M., Goriely, S. & Jorissen, A. Non-explosive hydrogen and helium burnings: abundance predictions from the NACRE reaction rate compilation. *Astron. Astrophys.* **347,** 572 (1999).
25. Amari, S. *et al.* Presolar grains from novae. *Astrophys. J.* **551,** 1065–1072 (2001).
26. Nittler, L. R. & Hoppe, P. Are presolar silicon carbide grains from novae actually from supernovae? *Astrophys. J.* **631,** L89–L92 (2005).
27. José, J. & Hernanz, M. The origin of presolar nova grains. *Meteorit. Planet. Sci.* **42,** 1135–1143 (2007).

Author Contributions T.K. wrote the text. T.K. and K.M.M. obtained and reduced the APEX data. N.A.P. obtained and reduced the SMA data. A.K. obtained and reduced the Effelsberg data. All authors contributed to the interpretation of the data and commented on the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.K. (tkaminsk@eso.org).

METHODS

APEX observations. CK Vul was observed with the APEX 12-m telescope[28] on several nights between 4 and 19 May 2014, and between 9 and 21 July 2014. Numerous frequency setups were observed between 217 GHz and 909 GHz, all of which are listed in Extended Data Table 2. For observations up to 270 GHz, we used the SHeFI/APEX-1 receiver[29] which operates in a single sideband mode and produces spectra in a 4-GHz-wide band. For frequencies between 278 GHz and 492 GHz, we used the FLASH$^+$ receiver[30] which operates simultaneously in two atmospheric bands at about 345 GHz and 460 GHz. Additionally, FLASH$^+$ separates the two heterodyne sidebands in the two 345/460 channels, giving four spectra simultaneously, each 4 GHz wide. Both APEX-1 and FLASH$^+$ are single-receptor receivers allowing for observation of one position at a time. For three of our setups with frequencies above 690 GHz, we used the CHAMP$^+$ receiver, which consists of two arrays operating in the atmospheric windows at 660 GHz and 850 GHz. Each CHAMP$^+$ array has seven receptors[31]. Each of the fourteen receptors of the CHAMP$^+$ array produced a single-sideband spectrum covering 2.8 GHz. As the backend (spectrometer) for the APEX-1 and FLASH$^+$ observations, we used the eXtended Fast Fourier Transform Spectrometer[32] (XFFTS) which provided us with a spectral resolution of 88.5 kHz. The CHAMP$^+$ spectra were acquired with the array version of FFTS which operates at the spectral resolution of 732 Hz.

For most of our spectral setups which cover CO transitions up to $J = 4–3$, we applied the position switching method with a reference at an offset $(-180'', -100'')$ from CK Vul, which was free of interstellar emission. Higher-J transitions of CO, all lines of ^{13}CO, and all setups which do not contain CO lines were observed with symmetric wobbler switching with a typical throw of $100''$.

Observations were performed in weather conditions that were excellent or optimal for the given frequency setup. The typical system temperatures (T_{sys}) and root mean square (r.m.s.) noise levels reached are given in Extended Data Table 2 (the r.m.s. is specified for spectral binning given in the sixth column of the table). The beam sizes and the main-beam efficiencies η_{mb} of the APEX antenna at each observed frequency is also given in the table. The typical calibration uncertainties are below 20%. All spectra were reduced using standard procedures in the CLASS/GILDAS package and converted to units of the main-beam brightness temperature (T_{mb}).

Effelsberg observations. The Effelsberg 100-m telescope was used to observe the classical circumstellar radio transitions: SiO(1–0) at $v = 1$ and 2; four ground-state transitions of OH $^2\Pi_{3/2}$ (between 1.6 GHz and 1.7 GHz); the $6_{1,6}$–$5_{2,3}$ transition of water at 22.235 GHz; and three lowest inversion lines of NH$_3$. From those, only the ammonia lines were detected and these observations are described in more detail below.

The three inversion lines of ammonia, $(J,K) = (1,1)$, $(2,2)$, and $(3,3)$ (the first two are para and the last is an ortho transition) we observed simultaneously on 2 August 2014. The secondary-focus receiver S13mm and the Effelsberg XFFTS were used. Spectra were centred at 23.750 GHz and covered 0.5 GHz at a resolution of 0.2 km s^{-1}. The spectra were moderately affected by baseline irregularities. The three lines of NH$_3$ are detected at a high signal-to-noise ratio (>10 for peaks), but baseline imperfections cast doubts on the actual profile and total intensity of the (3,3) line. The integration resulted in an r.m.s. noise level of 7.0 mK (in T_{mb} scale) per 9 km s^{-1} bin. The telescope beam had a full-width at half-maximum (FWHM) of $36.5''$.

Observations were repeated with the same instrumentation on 11 September 2014 but with the band centre shifted to lower frequencies to cover the (5,5) transition of ^{15}NH$_3$ at 23.42 GHz. The spectra covered the (1,1) and (2,2) lines of NH$_3$, but not the (3,3) transition. At the r.m.s. of 4.1 mK (T_{mb}) per 10 km s^{-1} the line of ^{15}NH$_3$ was not detected. This transition arises from a high level above the ground (with the energy of the upper level of $E_u = 296$ K) and may be very weak in this source.

SMA observations. To image the emission of selected lines discovered with APEX at a higher angular resolution, we used the SMA on 3 and 30 July 2014. On 3 July 2014, the array was used in its compact configuration and with eight operating antennas. The phase centre for all the SMA observations of CK Vul was the position of the radio continuum source measured by the Very Large Array[3] (VLA), that is, at right ascension (RA) = 19 h 47 min 38.074 s and declination (Dec.) = $+27°\,18'$ $45.16''$. As absolute-flux calibrators, MWC349a and Uranus were observed, while 3C279 and 3C454.3 were observed for a bandpass calibration; quasars 2025+337 and 2015+371 were our gain calibrators. The data covered four frequency ranges: 330.2–332.2 GHz, 335.2–337.2 GHz, 345.2–347.2 GHz, and 350.2–352.2 GHz. Although mainly aimed at observing the CO(3–2) transition, this setup gave us access to several emission lines and provided a very sensitive measurement of continuum emission. The system temperatures changed between 200 K and 500 K with the changing source elevation. The synthesized beam of these observations has a FWHM of $2.3'' \times 1.5''$ and a position angle (PA) of $87.6°$, while the primary beam, defining the field of view of the array, has an FWHM of $32''$.

On 30 July 2014, seven antennas were used in the subcompact configuration. The bandpass calibration was performed using observations of 3C279 while flux

calibration was obtained by observing Mars and Titan. Gain calibrators were the same as earlier. The typical system temperatures were between 90 and 130 K. We covered four frequency ranges: 216.9–218.8 GHz, 218.9–220.8 GHz, 228.9–230.8 GHz, and 230.9–232.8 GHz. The synthesized beam of these observations was $8.4'' \times 4.7''$ (PA = $71.5°$), while the primary beam at the observed frequencies is of $49''$.

The data were processed and calibrated in the MIR-IDL package (http://www.cfa.harvard.edu/sma/mir/). The calibrated visibilities were then imaged and further processed with Miriad[33]. The continuum emission was subtracted from the spectra as a best-fit first-order polynomial and continuum images were created by combining all four bands covering in total 8 GHz on each date. Resulting continuum flux densities are given in Extended Data Table 4.

A data inspection revealed that the interferometric maps of CO(3–2) show much lower flux than expected from the APEX spectra owing to the lack of short baselines. In the 345 GHz observation obtained in the compact configuration, the projected baselines gave us access to angular scales smaller than about $14''$. Any more extended emission was spatially filtered out by the interferometer. We corrected the interferometric observations by providing an APEX map covering a large part of the interferometer's field of view, that is, $11'' \times 11''$, and at a signal-to-noise ratio similar to that measured in the interferometric map. The two data sets were combined in Miriad using the *immerge* task.

Identification of spectral features. In the spectral survey obtained with APEX, we have identified 47 features to which we ascribed molecular transitions; three extra transitions were observed with the Effelsberg telescope. All lines are listed in Extended Data Table 2. Ten of these features are very weak so that their presence and/ or identification is uncertain. In the identification procedure, we referred to the Jet Propulsion Laboratory catalogue[34] and the Cologne Database for Molecular Spectroscopy[35,36] (CDMS). Extended Data Table 2 includes basic measurements for the strongest features: the centroid position with respect to the laboratory frequency of the ascribed transition; the line FWHM in velocity units; and profile-integrated intensity of the line in T_{mb} units. The list of detected transition includes mostly simple two-atomic species, but two molecules containing four atoms, that is, H$_2$CO and NH$_3$, were observed. Transitions of molecules containing H and CNO elements dominate the spectrum; those include CO, CN, HCN, HNC, HCO$^+$, N$_2$H$^+$, H$_2$CO (and their isotopologues). The strongest are lines of carbon monoxide. Our survey covered four transitions of the main CO species and at least three transitions of its rare isotopologues. Only two unambiguously identified molecules are carriers of heavier atoms, that is, SiO and CS, the latter being identified only tentatively. Two ionic species have been firmly identified, HCO$^+$ and N$_2$H$^+$. The most striking feature of the list of detected transitions is the high number of lines from rare isotopologues of CNO elements.

Determination of abundances and excitation temperatures. A few molecules were observed in multiple transitions within a range of E_u wide enough to allow a simple excitation analysis. With the aim of constraining the excitation temperatures and column densities, we performed analysis of rotational diagrams[37], in which we assumed thermodynamic equilibrium, optically thin emission, and that the gas is isothermal. Although some of the observed transitions are likely to be optically thick and the gas is not isothermal, this initial analysis was aimed to get the first constraints on the gas physical parameters. We used least-squares fitting to derive the physical parameters. Partition functions were interpolated from data tabulated in CDMS. The sizes of the emission regions, necessary for a beam-filling correction, were based on our interferometric maps.

Our rotational diagram analysis was supported by spectra simulations performed in CASSIS[38]. The tool allowed us to generate a model spectrum with line profiles approximated by Gaussians. The simulation was based on the same assumptions as underlying the rotational diagram analysis, but included a limited correction for line saturation effects. The CASSIS simulation was especially helpful in an analysis of blended features and transitions with considerable hyperfine splitting, for instance CN and its isotopologues.

Rotational diagrams for CO and H^{13}CN, which were also observed in transitions with $E_u > 80$ K, cannot be reproduced by a simple linear fit. This is probably a consequence of multiple gas components at different temperatures (or a continuous range of temperatures), combined with different sizes of the emission regions contributing most to the given transition. Additionally, those transitions at high E_u were typically observed at high frequencies at which the APEX beam is much smaller than for the rest of the observed transitions and does not encompass the entire molecular region. Because of the missing spatial information, those transitions were omitted in the rotational diagram analysis. Here we focus on the gas at lower temperatures, which dominates the emission in lower rotational transitions.

For most molecules analysed here, the excitation temperature was derived from the rotational diagram of the isotopologue for which the highest number of transitions was observed. Then, the same temperature was assumed for other isotopologues, and column densities were calculated for all other isotopic species observed in at least one transition. For all three CO isotopologues, good temperature estimates

were obtained for each isotopologue and the final column densities were calculated for a weighted mean of the three values. While the relative abundances of the different species analysed here are subject to large errors (mainly because of the complex spatio-kinematical structure of the gas), the isotopic ratios are much more reliable—they weakly depend on the temperature and are not directly sensitive to the details of the spatial distribution (if no chemical fractionation takes place). They are, however, affected by opacity effects (see below). In Table 1, we therefore report only the isotopic ratios resulting from our analysis. To put constraints on species containing the oxygen isotope ^{17}O, we used the upper limit on the $C^{17}O(3–2)$ line covered by APEX.

The $^{12}C/^{13}C$ ratio was derived for four species and is consistently found to be 2–3 for three of them. The value derived from CO line ratios is an outlier, with a slightly higher ratio of about 6. The saturation effect, if present, should be strongest in the CO transitions, giving a ratio that is lower than in the weaker lines of the rarer species. The nitrogen-bearing species, HCN and CN, lead to two different values of the isotopic $^{14}N/^{15}N$ ratio, 26 and 4, respectively. There is an extra uncertainty in the abundance analysis of the weak $C^{15}N$ spectra related to their hyperfine structure and blending. Chemical fractionation cannot be excluded because CN is probably a product of photodissociation of HCN and self-shielding effects are likely to occur for HCN isotopologues.

The rotational diagram analysis allowed us to derive excitation temperatures for the different species. They are typically in the range 8–22 K, but extra gas components at higher temperatures are evident in transitions from higher energy levels.

We tried to assess the influence of the line saturation effects on the results of our analysis by investigating the optical depth of the CO lines, which are expected to have the highest opacity. We analysed the CO emission over the full line profile (-220 km s^{-1} to $+200$ km s^{-1}, where the velocity is expressed with respect to the local standard of rest, V_{LSR}) and also in one wing (-50 km s^{-1} to $+40$ km s^{-1}). The opacity was calculated for the best-fitting parameters of temperature and column density. The line FWHM was set to 120 km s^{-1} and 90 km s^{-1} for the full profile and the probed part of the wing, respectively. For a source size of 10″, whose solid angle is equivalent to that of the entire emission region seen in the combined SMA and APEX maps, we get an optical thickness of $\tau_0 = 0.35$ for CO(2–1) (strongest feature observed) and lower values for the weaker lines. For the emission in the wing, we get $\tau_0 = 0.18$ for the $J = 2–1$ transition and much less for the higher-J lines. However, the obtained results are sensitive to the adopted value of the source size. For FWHM = 6.8″, which corresponds to the size of the CO(3–2) emission at the isophote at the 30% of the peak, the strongest line would have an optical thickness of 0.75. Then the central CO component, which is of an even smaller size of 1″–2″ and contributes about 15% of the total observed flux, produces emission of moderate opacity of the order of 1. Only if the emission arises in compact clumps are the lines optically thick.

APEX observations of other Galactic novae. Our detection of CO in CK Vul contradicts the previous claims of non-detection of rotational circumstellar lines in this source[39]. It also casts doubts on all earlier negative results of searches of submillimetre-wave lines towards novae and related objects. Observation of lines as broad as those expected in novae (300–7,000 km s^{-1}) are very demanding in terms of the atmospheric and instrumental stability. In the earlier attempts, lines were often broader than the full available spectral range of the receiving systems or comparable in width to typical baseline ripples. The presence of molecular emission in classical novae was therefore tested anew using the modern instrumentation of APEX.

The novae observed with APEX were selected from ref. 40 using the following selection criteria: (1) the source has to reach elevations higher than 40°, and (2) it has to be available for observations in the local sidereal time range 23–13 h to not collide with the inner-Galaxy projects in the APEX observing queue, (3) it must be located at least 3° from the Galactic plane to avoid contamination from Galactic CO emission. These requirements limited the number of sources to 17, which are listed in Extended Data Table 2.

The observations were performed between 24 and 28 August 2014 and on 8 September 2014 with FLASH$^+$ connected to FFTS providing a spectral coverage of 4 GHz. Although four spectral ranges were covered simultaneously, the observing procedure was optimized for the band centred at the frequency of the CO(3–2) line. The observations were performed with wobbler switching with a throw of 80″. No source was detected in the CO(3–2) line at the typical r.m.s. of 2.5 mK (T_{mb}) per 33 km s^{-1} (Extended Data Table 3). At the same sensitivity the line was very clearly seen in the spectrum of CK Vul.

Spectral energy distribution. Using archival and literature data combined with our SMA continuum measurements, we constructed the spectral energy distribution of CK Vul. The data are described in detail at the end of this section; the measurements are summarized in Extended Data Table 4 and shown in Extended Data Fig. 1.

The spectral energy distribution is dominated by emission ranging from about 20 μm up to the millimetre wavelengths. The flux density F_ν peaks at about 100 μm. The long-wavelength part of the F_ν distribution, from the far-infrared to the SMA

measurement, has a slope with a spectral index $\alpha = 2.1 \pm 0.1$ (where $F_\nu \propto \nu^\alpha$ and ν is the frequency) and can be interpreted as thermal dust radiation. A single blackbody cannot explain the observed emission entirely but the best fit of a single Planck function provides a rough estimate of the dust temperature of 39 ± 5 K. The best fit of a greybody, that is, a Planck function multiplied by dust emissivity in the form of a power law ν^β, gives a temperature of 15 K and $\beta = 1.0$. This fit underestimates the source fluxes at shorter wavelengths, but we believe it provides a good estimate on the value of β. Moreover, $\beta \cong 1.0$ is expected for circumstellar dust in the form of amorphous carbon or layer-lattice silicates[41]; $\beta \approx 1.0$ is also typical for circumstellar disks[42,43]. We note that the chemical composition and the form (crystalline/amorphous) of dust in CK Vul remains completely unknown. To better reproduce the flux at short wavelengths, we also obtained a fit of two grey bodies with β being fixed at a value of 1.0. This gave temperatures of 15 K and 49 K. It is unlikely that the dust is characterized by two isothermal components. Instead, one can expect a continuous range of temperatures in 15–49 K. The fit of two grey bodies (Extended Data Fig. 1) underestimates the fluxes around 160 μm. Although this could be overcome by introducing an extra component at an intermediate temperature, we did not attempt it because the least-square fits become degenerate at the required number of parameters.

The flux under the reconstructed spectral energy distribution is 6.0×10^{-11} erg s^{-1} cm^{-2}. Adopting the distance of 700 pc (ref. 3), we calculate the source luminosity to be 3.6×10^{33} erg s^{-1} (or 0.9 solar luminosities). This luminosity is close to the 0.7 solar luminosities found from ionization-equilibrium calculations for the optical nebula[3] (here corrected to the distance of 700 pc). The dust emission we observe must be reprocessed radiation of the central source which is hidden for our line of sight at wavelengths shorter than ~20 μm. Because the obscuring material has a form of a flattened, torus-like structure, the radiation field within the whole system is anisotropic. Our estimate should therefore be treated as a lower limit on the actual luminosity of the source.

Continuum observations and data reductions. *Herschel.* On 23 October 2011, CK Vul was serendipitously observed by photometers on board the Herschel Space Observatory in a field covered within the Hi-Gal project[44]. Two scans (OBSIDs 1342231339 and 1342231340) were obtained in orthogonal directions across a large field covering CK Vul. The two Herschel cameras, PACS and SPIRE, were used simultaneously in these observations. In both scans, PACS was used with its blue (70 μm) and red (160 μm) bands (that is, the green band was not used) and SPIRE produced maps in its all three bands, that is, 250 μm, 350 μm, and 500 μm. Data were retrieved from the Herschel Science Archive and processed in the Herschel Interactive Processing Environment (HIPE). The raw data were automatically reduced by the standard pipeline which used the calibration scheme version 12.1. The pointing accuracy of Herschel is typically 2″ and the source we identify as CK Vul has a position that is consistent within 3″ with the position of the continuum seen by the VLA and SMA. In all the observed bands, CK Vul appears as a point source, but its background becomes more and more contaminated by diffuse Galactic emission with increasing wavelength. A bright source closest to CK Vul is located 1.5 arcmin west. It is weaker than CK Vul in all the PACS and SPIRE bands.

Source fluxes in the four individual PACS maps were measured with aperture-photometry techniques including background subtraction and a correction for limited aperture size. Results obtained for the two PACS bands were averaged and the standard deviation from the two measurements in each band was taken as the uncertainty.

Source fluxes in the SPIRE observations were measured using aperture photometry tasks and a 'timeline fitting' procedure available in HIPE. In addition to an aperture correction, we also applied a colour correction to the measured fluxes using tabular data included in the Herschel-SPIRE calibration data for the spectral index of $\alpha = 1.0$. Measurements were obtained on individual scans and the results were averaged for the given band. The uncertainties in the absolute flux calibration are 6% for SPIRE, and 10% and 20% for the blue and red bands of PACS, respectively.

Spitzer. CK Vul was observed multiple times with Spitzer instruments. The Multiband Imaging Photometer for Spitzer (MIPS) operating in bands at 24 μm, 70 μm, and 160 μm observed the position on two different dates. On 17 October 2004 a MIPS scan centred on the object was obtained (AOR 10837504, Principal Investigator (PI) A. Evans) and on 7 October 2005 the position was covered by a scan aiming to observe Galactic emission in the field of CK Vul (AOR 15621888, PI S. Carey; no data in the 160 μm band were collected). We used the pipeline processed data and aperture-photometry procedures to derive the source fluxes. The aperture- and colour-corrected (for the assumed blackbody spectrum of 30 K) fluxes are listed in Extended Data Table 4. For the 24 μm and 70 μm bands we list the average flux from the two scans and the standard deviation from the mean as an error. The single observation in the 160 μm band was spatially under-sampled and only a very rough flux estimate was performed. The flux is indeed lower than the PACS measurement at similar wavelengths and was omitted in the analysis. The measurement at 70 μm, on the other hand, agrees very well with that from PACS at a similar wavelength. At

the angular resolution of the MPIS maps at 24 μm and 70 μm of 6″ and 18″ (FWHM), respectively, the source appears point-like.

The InfraRed Array Camera (IRAC) observed the positions of CK Vul four times in October 2004 and December 2012 with a different combination of IRAC bands (3.6 μm, 5.8 μm, 4.5 μm, and 8.0 μm). None of the IRAC maps shows a measurable source at the position of CK Vul. We used the most sensitive scans in the 3.6 μm and 8.0 μm bands to derive upper limits on the emission from CK Vul. The standard deviation of the flux at the position of the object is of about $\sigma = 1.21\,\mu$Jy and 4.24 μJy in the 3.6 μm and 8.0 μm bands, respectively.

WISE. The point source catalogue of the Wide-field Infrared Survey Explorer (WISE) survey lists a source consistent with the position of CK Vul, which was measured in three out of the four WISE bands (there is only an upper limit in the W3 band). The source catalogue position is about 5″ away from the SMA position of the continuum source. The catalogue flags also indicate that the source is resolved (FWHM of the point-spread functions are 6.1″, 6.4″, 6.5″, and 12″ in the W1 to W4 bands.) The flags also indicate that the source is variable in the W1 and W2 bands. After inspecting the WISE images covering the position of CK Vul and comparing them to optical and radio maps, we concluded that only the W4 measurement at 22 μm can be definitely ascribed to the source seen at longer wavelengths (while the W1 and W2 data correspond to 'variable 2' identified in a recent study[16]). The average magnitudes from the WISE point source catalogue (resulting from profile fitting) were converted to flux units using standard zero points[45] and are listed in Extended Data Table 4. The catalogue values in W1 to W3 bands can all be treated here as rough upper limits on the flux of CK Vul. No colour correction was applied.

AKARI. Point source catalogues[46] of the AKARI satellite mission contain one source that matches the positions of CK Vul. The measurements obtained with the Far-Infrared Surveyor (FIS) instrument, which operates at 65 μm, 90 μm, 140 μm, and 160 μm are flagged as reliable only for the measurement at 90 μm (722 mJy). In the source catalogue of the Infrared Camera (IRC) survey at 9 μm and 18 μm, no source can be identified as CK Vul.

JCMT. Literature data[47] exist based on observations obtained with the James Clerk Maxwell Telescope (JCMT) and the SCUBA bolometer at about 450 μm and 850 μm (~667 GHz and ~353 GHz). The measurement at 850 μm covers a wavelength range close to that of one of our SMA observations. The SCUBA flux is slightly above that derived in the SMA observations. The reason for this is probably the fact that our SMA measurements represent only line-free continuum while the bolometric observations represent summary flux of continuum and emission lines. Our APEX spectra in the range between 333 GHz and 357 GHz, which overlap with a high-sensitivity part of the SCUBA 850 μm bandpass, show a line flux density of 94.3 mJy, which constitutes 42% of the flux measured with SCUBA. Spectral lines contribute substantially therefore to the bolometric measurements, at least in the submillimetre-wave region. To a lesser degree, the SMA continuum measurement at 341 GHz can be partially lower than that measured with JCMT because extended continuum emission, if present, was partially filtered out by the interferometer.

The SCUBA measurement at 450 μm is close in wavelength to the SPIRE 500 μm band (482.3 μm), but has a much lower flux. Compared to all the data collected, this SCUBA measurement is a clear outlier. Because the ground-based observations at 450 μm are very demanding in terms of weather conditions, we suspect that this measurement has an extra systematic uncertainty not quoted in the work reporting the data[47].

On 3 August 2012 CK Vul was observed again with the JCMT, this time with the SCUBA-2 bolometer array. While no source was detected at 450 μm, the emission at 850 μm is very clear. Using the archival pipeline-processed data, we measured the source flux to be 194.0 ± 1.7 mJy (1σ error). This measurement agrees within the uncertainties with the flux measured in observations taken eleven years earlier[47]. The source is unresolved at the resolution of 14.5″ (FWHM). The 3σ upper limit on the flux density at 450 μm is 1.08 Jy.

VLA and Toruń. For completeness, in the spectral energy distribution analysis we include flux measurements of the radio continuum obtained with the VLA[3,48] and the OCRA-p receiver at the Toruń radio telescope[49].

28. Güsten, R. *et al.* The Atacama Pathfinder EXperiment (APEX)—a new submillimeter facility for southern skies. *Astron. Astrophys.* **454,** L13–L16 (2006).
29. Vassilev, V. *et al.* A Swedish heterodyne facility instrument for the APEX telescope. *Astron. Astrophys.* **490,** 1157–1163 (2008).
30. Klein, T. *et al.* FLASH⁺—a dual-channel wide-band spectrometer for APEX. *IEEE Trans. Terahertz Sci. Technol.* **4,** 588–596 (2014).
31. Kasemann, C. *et al.* CHAMP⁺: a powerful array receiver for APEX. *SPIE Conf. Ser.* **6275,** 62750N (2006).
32. Klein, B. *et al.* High-resolution wide-band fast Fourier transform spectrometers. *Astron. Astrophys.* **542,** L3 (2012).
33. Sault, R. J., Teuben, P. J. & Wright, M. C. H. A retrospective view of MIRIAD. *Astron. Data Analysis Software Syst. IV* **77,** 433 (1995).
34. Pickett, H. M. *et al.* Submillimeter, millimeter and microwave spectral line catalog. *J. Quant. Spectrosc. Radiat. Transf.* **60,** 883–890 (1998).
35. Müller, H. S. P., Schlöder, F., Stutzki, J. & Winnewisser, G. The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *J. Mol. Struct.* **742,** 215–227 (2005).
36. Müller, H. S. P. *et al.* The Cologne Database for Molecular Spectroscopy, CDMS: a tool for astrochemists and astrophysicists. *IAU Symp.* **235,** 62P (2005).
37. Goldsmith, P. F. & Langer, W. D. Population diagram analysis of molecular line emission. *Astrophys. J.* **517,** 209–225 (1999).
38. Vastel, C. *Formalism for the CASSIS Software* http://cassis.irap.omp.eu/docs/RadiativeTransfer.pdf (2014).
39. Weight, A., Evans, A., Albinson, J. S. & Krautter, J. Millimetre observations of old novae. *Astron. Astrophys.* **268,** 294–298 (1993).
40. Williams, G. *CBAT List of Novae in the Milky Way* http://www.cbat.eps.harvard.edu/nova_list.html (2014).
41. Tielens, A. G. G. M. & Allamandola, L. J. Composition, structure, and chemistry of interstellar dust. *Interstellar Processes* **134,** 397–470 (1987).
42. Beckwith, S. V. W. & Sargent, A. I. Particle emissivity in circumstellar disks. *Astrophys. J.* **381,** 250–258 (1991).
43. Draine, B. T. On the submillimeter opacity of protoplanetary disks. *Astrophys. J.* **636,** 1114–1120 (2006).
44. Molinari, S. *et al.* Hi-GAL: The Herschel Infrared Galactic Plane Survey. *Publ. Astron. Soc. Pacif.* **122,** http://dx.doi.org/101086/651314 (2010).
45. Wright, E. L. *et al.* The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance. *Astron. J.* **140,** 1868–1881 (2010).
46. Yamauchi, C. *et al.* AKARI-CAS—online service for AKARI all-sky catalogues. *Publ. Astron. Soc. Pacif.* **123,** 852–864 (2011).
47. Evans, A. *et al.* CK Vul: reborn perhaps, but not hibernating. *Mon. Not. R. Astron. Soc.* **332,** L35–L38 (2002).
48. Bode, M. F., Seaquist, E. R. & Evans, A. Radio survey of classical novae. *Mon. Not. R. Astron. Soc.* **228,** 217–227 (1987).
49. Hajduk, M. *et al.* On the evolved nature of CK Vul. *ASP Conf. Ser.* **391,** 151 (2008).

LETTER

# An ultrafast rechargeable aluminium–ion battery

Meng-Chang Lin[1,2]*, Ming Gong[1]*, Bingan Lu[1,3]*, Yingpeng Wu[1]*, Di-Yan Wang[1,4,5], Mingyun Guan[1], Michael Angell[1], Changxin Chen[1], Jiang Yang[1], Bing-Joe Hwang[6] & Hongjie Dai[1]

The development of new rechargeable battery systems could fuel various energy applications, from personal electronics to grid storage[1,2]. Rechargeable aluminium-based batteries offer the possibilities of low cost and low flammability, together with three-electron-redox properties leading to high capacity[3]. However, research efforts over the past 30 years have encountered numerous problems, such as cathode material disintegration[4], low cell discharge voltage (about 0.55 volts; ref. 5), capacitive behaviour without discharge voltage plateaus (1.1–0.2 volts[6] or 1.8–0.8 volts[7]) and insufficient cycle life (less than 100 cycles) with rapid capacity decay (by 26–85 per cent over 100 cycles)[4–7]. Here we present a rechargeable aluminium battery with high-rate capability that uses an aluminium metal anode and a three-dimensional graphitic-foam cathode. The battery operates through the electrochemical deposition and dissolution of aluminium at the anode, and intercalation/de-intercalation of chloroaluminate anions in the graphite, using a non-flammable ionic liquid electrolyte. The cell exhibits well-defined discharge voltage plateaus near 2 volts, a specific capacity of about 70 mA h g^{-1} and a Coulombic efficiency of approximately 98 per cent. The cathode was found to enable fast anion diffusion and intercalation, affording charging times of around one minute with a current density of ~4,000 mA g^{-1} (equivalent to ~3,000 W kg^{-1}), and to withstand more than 7,500 cycles without capacity decay.

Owing to the low-cost, low-flammability and three-electron redox properties of aluminium (Al), rechargeable Al-based batteries could in principle offer cost-effectiveness, high capacity and safety, which would lead to a substantial advance in energy storage technology[3,8]. However, research into rechargeable Al batteries over the past 30 years has failed to compete with research in other battery systems. This has been due to problems such as cathode material disintegration[4], low cell discharge voltage (~0.55 V; ref. 5), capacitive behaviour without discharge voltage plateaus (1.1–0.2 V, or 1.8–0.8 V; refs 6 and 7, respectively), and insufficient cycle life (<100 cycles) with rapid capacity decay (by 26–85% over 100 cycles)[4–7]. Here we report novel graphitic cathode materials that afford unprecedented discharge voltage profiles, cycling stabilities and rate capabilities for Al batteries.

We constructed Al/graphite cells (see diagram in Fig. 1a) in Swagelok or pouch cells, using an aluminium foil (thickness ~15–250 μm) anode, a graphitic cathode, and an ionic liquid electrolyte made from vacuum dried AlCl$_3$/1-ethyl-3-methylimidazolium chloride ([EMIm]Cl; see Methods, residual water ~500 p.p.m.). The cathode was made from either pyrolytic graphite (PG) foil (~17 μm) or a three-dimensional graphitic foam[9,10]. Both the PG foil and the graphitic-foam materials exhibited typical graphite structure, with a sharp (002) X-ray diffraction (XRD) graphite peak at $2\theta \approx 26.55°$ (d spacing, 3.35 Å; Extended Data Fig. 1). The cell was first optimized in a Swagelok cell operating at 25 °C with a PG foil cathode. The optimal ratio of AlCl$_3$/[EMIm]Cl was found to be ~1.3–1.5 (Extended Data Fig. 2a), affording a specific discharging capacity of 60–66 mA h g^{-1} (based on graphitic cathode mass) with a Coulombic efficiency of 95–98%. Raman spectroscopy revealed that with an AlCl$_3$/[EMIm]Cl ratio of ~1.3, both AlCl$_4^-$ and Al$_2$Cl$_7^-$ anions were present (Extended Data Fig. 2b) at a ratio [AlCl$_4^-$]/[Al$_2$Cl$_7^-$] ≈ 2.33
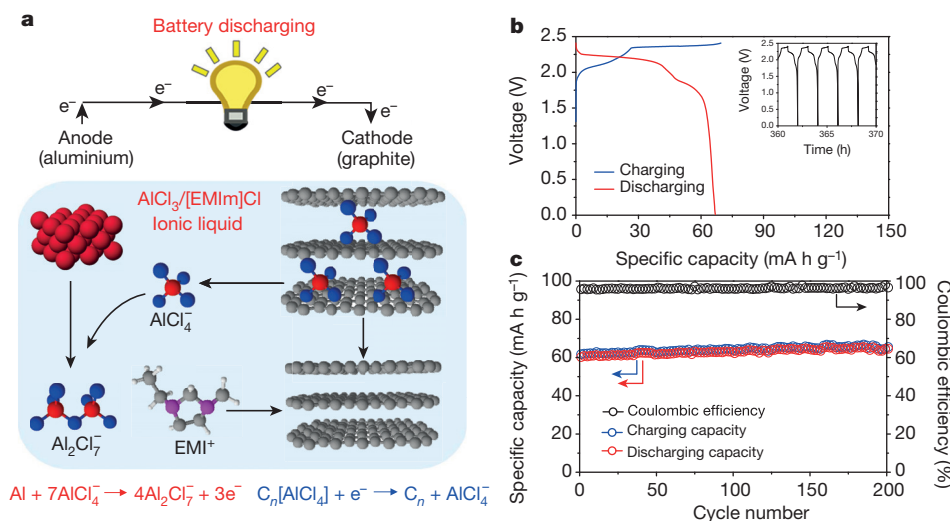
Figure 1 | Rechargeable Al/graphite cell. a, Schematic drawing of the Al/graphite cell during discharge, using the optimal composition of the AlCl$_3$/[EMIm]Cl ionic liquid electrolyte. On the anode side, metallic Al and AlCl$_4^-$ were transformed into Al$_2$Cl$_7^-$ during discharging, and the reverse reaction took place during charging. On the cathode side, predominantly AlCl$_4^-$ was intercalated and de-intercalated between graphite layers during charge and discharge reactions, respectively. **b**, Galvanostatic charge and discharge curves of an Al/pyrolytic graphite (PG) Swagelok cell at a current density of 66 mA g^{-1}. Inset, charge and discharge cycles. **c**, Long-term stability test of an Al/PG cell at 66 mA g^{-1}.

Reactions in figure:
Al + 7AlCl$_4^-$ → 4Al$_2$Cl$_7^-$ + 3e$^-$ C$_n$[AlCl$_4$] + e$^-$ → C$_n$ + AlCl$_4^-$

[1]Department of Chemistry, Stanford University, Stanford, California 94305, USA. [2]Green Energy and Environment Research Laboratories, Industrial Technology Research Institute, Hsinchu 31040, Taiwan. [3]School of Physics and Electronics, Hunan University, Changsha 410082, China. [4]Department of Chemistry, National Taiwan Normal University, Taipei 11677, Taiwan. [5]Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei 10617, Taiwan. [6]Department of Chemical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan.
*These authors contributed equally to this work.

(ref. 11). The cathode specific discharging capacity was found to be independent of graphite mass (Extended Data Fig. 3), suggesting that the entirety of the graphite foil participated in the cathode reaction.
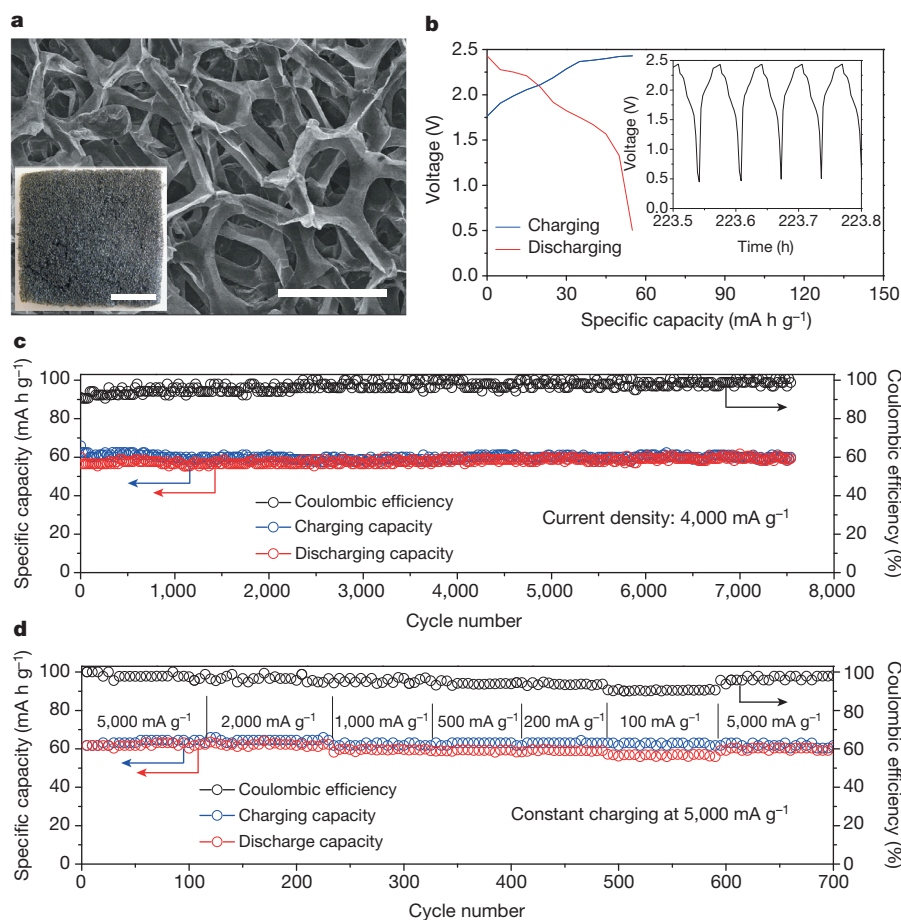
The Al/PG cell exhibited clear discharge voltage plateaus in the ranges 2.25–2.0 V and 1.9–1.5 V (Fig. 1b). The relatively high discharge voltage plateaus are unprecedented among all past Al-ion charge-storage systems[4–7]. Similar cell operation was observed with the amount of electrolyte lowered to ~0.02 ml per mg of cathode material (Extended Data Fig. 4). Charge–discharge cycling at a current density of 66 mA g^{-1} (1 C charging rate) demonstrated the high stability of the Al/PG cell, which nearly perfectly maintained its specific capacity over >200 cycles with a 98.1 ± 0.4% Coulombic efficiency (Fig. 1c). This was consistent with the high reversibility of Al dissolution/deposition, with Coulombic efficiencies of 98.6–99.8% in ionic liquid electrolytes[12–15]. No dendrite formation was observed on the Al electrode after cycling (Extended Data Fig. 5). To maintain a Coulombic efficiency >96%, the cut-off voltage of the Al/PG cell (that is, the voltage at which charging was stopped) was set at 2.45 V, above which reduced efficiencies were observed (see Extended Data Fig. 6a), probably due to side reactions (especially above ~2.6 V) involving the electrolyte, as probed by cyclic voltammetry with a glassy carbon electrode against Al (Extended Data Fig. 6b).

We observed lowered Coulombic efficiency and cycling stability of the Al/graphite cell when using electrolytes with higher water contents, up to ~7,500 p.p.m. (Extended data Fig. 6c, d), accompanied by obvious H$_2$ gas evolution measured by gas chromatography (Extended Data Fig. 6e). This suggested side reactions triggered by the presence of residual water in the electrolyte, with H$_2$ evolution under reducing potential on the Al side during charging. Further lowering the water content

of the ionic liquid electrolyte could be important when maximizing the Coulombic efficiency of the Al/graphite cells.

The Al/PG cell showed limited rate capability with much lower specific capacity when charged and discharged at a rate higher than 1 C (Extended Data Fig. 7). It was determined that cathode reactions in the Al/PG cell involve intercalation and de-intercalation of relatively large chloroaluminate (Al$_x$Cl$_y$$^{-}$) anions in the graphite (see below for XRD evidence of intercalation), and the rate capability is limited by slow diffusion of anions through the graphitic layers[16]. When PG was replaced by natural graphite, intercalation was evident during charging owing to dramatic expansion (~50-fold) of the cathode into loosely stacked flakes visible to the naked eye (Extended Data Fig. 8a). In contrast, expansion of PG foil upon charging the Al/PG cell was not observable by eye (Extended Data Fig. 8b), despite the similar specific charging capacity of the two materials (Extended Data Fig. 8c). This superior structural integrity of PG over natural graphite during charging was attributed to the existence of covalent bonding between adjacent graphene sheets in PG[17], which was not present in natural graphite. Using PG, which has an open, three-dimensionally-bound graphitic structure, we prevented excessive electrode expansion that would lead to electrode disintegration, while maintaining the efficient anion intercalation necessary for high performance.

Because high-rate and high-power batteries are highly desirable for applications such as electrical grid storage, the next step in the investigation was to develop a cathode material that would have reduced energetic barriers to intercalation during charging[16]. We investigated a flexible graphitic foam (Fig. 2a), which was made on a nickel foam template by chemical vapour deposition[9,10] (see Methods), as a possible material for


Figure 2 | An ultrafast and stable rechargeable Al/graphite cell. a, A scanning electron microscopy image showing a graphitic foam with an open frame structure; scale bar, 300 μm. Inset, photograph of graphitic foam; scale bar, 1 cm. **b,** Galvanostatic charge and discharge curves of an Al/graphitic-foam pouch cell at

a current density of 4,000 mA g^{-1}. **c,** Long-term stability test of an Al/graphitic-foam pouch cell over 7,500 charging and discharging cycles at a current density of 4,000 mA g^{-1}. **d,** An Al/graphitic-foam pouch cell charging at 5,000 mA g^{-1} and discharging at current densities ranging from 100 to 5,000 mA g^{-1}.

ultrafast Al batteries. The graphite whiskers in the foam were 100 μm in width (Fig. 2a), with large spaces in between, which greatly decreased the diffusion length for the intercalating electrolyte anions and facilitated more rapid battery operation.
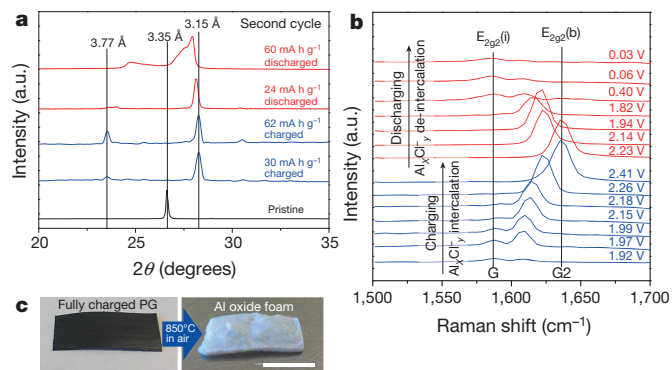
Remarkably, the Al/graphitic-foam cell (in a pouch cell configuration) could be charged and discharged at a current density up to 5,000 mA g^{-1}, about 75 times higher (that is, at a 75 C rate, <1 min charge/discharge time) than the Al/PG cell while maintaining a similar voltage profile and discharge capacity (∼60 mA h g^{-1}) (Figs 1b and 2b). An impressive cycling stability with ∼100% capacity retention was observed over 7,500 cycles with a Coulombic efficiency of 97 ± 2.3% (Fig. 2c). This is the first time an ultrafast Al-ion battery has been constructed with stability over thousands of cycles. The Al/graphitic-foam cell retained similar capacity and excellent cycling stability over a range of charge–discharge rates (1,000–6,000 mA g^{-1}) with 85−99% Coulombic efficiency (Extended Data Fig. 9a). It was also found that this cell could be rapidly charged (at 5,000 mA g^{-1}, in ∼1 min) and gradually discharged (down to 100 mA g^{-1}, Fig. 2d and Extended Data Fig. 9b) over ∼34 min while maintaining a high capacity (∼60 mA h g^{-1}). Such a rapid charging/variable discharging rate could be appealing in many real-world applications.

We propose that simplified Al/graphite cell redox reactions during charging and discharging can be written as:

$$4Al_2Cl_7^- + 3e^- \rightleftharpoons Al + 7AlCl_4^- \qquad (1)$$

$$C_n + AlCl_4^- \rightleftharpoons C_n[AlCl_4] + e^- \qquad (2)$$
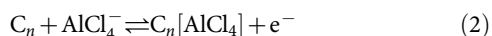
where n is the molar ratio of carbon atoms to intercalated anions in the graphite. The balanced $AlCl_4^-$ and $Al_2Cl_7^-$ concentrations in the electrolyte allowed for an optimal charging capacity at the cathode, with abundant $AlCl_4^-$ for charging/intercalation in graphite (equation (2)), and sufficient $Al_2Cl_7^-$ concentration for charging/electrodeposition at the anode (equation (1)).

Ex situ XRD measurement of graphite foil (Fig. 3a) confirmed graphite intercalation/de-intercalation by chloroaluminate anions during charging/discharging. The sharp pristine graphite foil (002) peak at $2\theta = 26.55°$ (d spacing = 3.35 Å) (Fig. 3a) vanished on charging to a specific capacity of ∼30 mA h g^{-1}, while two new peaks appeared at ∼28.25° ($d ≈ 3.15$ Å) and ∼23.56° ($d ≈ 3.77$ Å) (Fig. 3a), with peak intensities further increasing on fully charging to ∼62 mA h g^{-1}. The doublet XRD peak suggested highly strained graphene stacks formed on anion intercalation[18]. Analysis of the peak separation (see Methods) suggested a stage 4 graphite intercalation compound with an intercalant gallery height (spacing between adjacent graphitic host layers) of
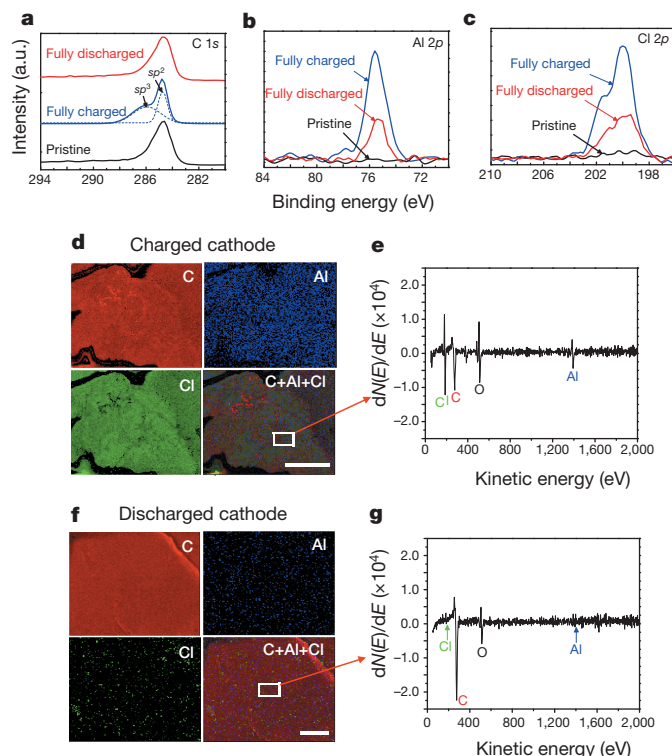
∼5.7 Å, indicating that the $AlCl_4^-$ anions (size ∼5.28 Å; ref. 19) were intercalated between graphene layers in a distorted state. Full discharging led to the recovery of the graphite peak but with a broad shoulder (Fig. 3a), probably caused by irreversible changes in the stacking between the graphene layers or a small amount of trapped species.

In situ Raman spectroscopy was also performed to probe chloroaluminate anion intercalation/de-intercalation from graphite during cell charge/discharge (Fig. 3b). The graphite G band (∼1,584 cm^{-1}) diminished and split into a doublet (1,587 cm^{-1} for the E_{2g2}(i) mode and ∼1,608 cm^{-1} for the E_{2g2}(b) mode) upon anion intercalation (Fig. 3b)[20], and then evolved into a sharp new peak (∼1,636 cm^{-1}, the G2 band of the E_{2g2}(b) mode, spectrum 2.41 V, Fig. 3b) once fully charged. The spectral changes were then reversed upon discharging (Fig. 3b), as the typical graphite Raman G band (1,584 cm^{-1}) was recovered when fully discharged (spectrum 0.03 V, Fig. 3b). Similar Raman spectra and XRD data were obtained with a graphitic-foam cathode (Extended Data Fig. 10a, b). Interestingly, calcination of a fully charged PG foil at 850 °C in air (Fig. 3c) yielded a white aluminium oxide foam (Extended Data Fig. 10c), confirming the intercalation of chloroaluminate anions into the carbon network, which had been evidently removed oxidatively.

Lastly, X-ray photoelectron spectra (XPS) and Auger electron spectroscopy (AES) were performed to probe the chemical nature of the intercalated species in our graphitic cathodes (see Methods for details). To minimize the amount of trapped electrolyte, graphitic foam was used and the electrode was thoroughly washed with anhydrous methanol. XPS revealed that upon charging pristine graphite, the 284.8 eV C 1s peak developed a shoulder at higher energy (∼285.9 eV, Fig. 4a), confirming electrochemical oxidation of graphitic carbon by intercalation of $AlCl_4^-$ anions (equation (2)). Chloroaluminate intercalation was evident from the appearance of Al 2p and Cl 2p peaks (Fig. 4b, c). Upon

Figure 3 | Al/graphite cell reaction mechanisms. a, *Ex situ* X-ray diffraction patterns of PG in various charging and discharging states through the second cycle. **b,** *In situ* Raman spectra recorded for the PG cathode through a charge–discharge cycle, showing chloroaluminate anion intercalation/de-intercalation into graphite. **c,** After calcination of a fully charged (62 mA h g^{-1}) PG electrode at 850 °C in air, the sample completely transformed into a white foam made of aluminium oxide. Scale bar, 1 cm.

Figure 4 | Chemical probing of a graphitic cathode by XPS and AES. a, XPS data of the C 1s peak of a graphitic-foam electrode: pristine, fully charged and fully discharged. **b, c,** XPS data of Al 2p and Cl 2p peaks observed with a graphitic-foam electrode: pristine, fully charged and fully discharged. **d–g,** AES mapping images for C, Al and Cl (**d, f**), and the AES spectrum of the boxed regions (**e, g**) obtained with a fully charged graphitic-foam sample (**d, e**) and a fully discharged graphitic-foam sample (**f, g**). Scale bars: **d,** 25 μm; **f,** 10 μm.

discharging, the C 1*s* XPS spectrum of the cathode reverted to that of the pristine graphite due to anion de-intercalation and carbon reduction (Fig. 4a). Also, a substantial reduction in the Al 2*p* and Cl 2*p* signals was recorded over the graphite sample (see Fig. 4b, c). The remaining Al and Cl signals observed were attributed to trapped/adsorbed species in the graphite sample, which was probed by XPS over a large area. Furthermore, high spatial resolution AES elemental mapping of a single graphite whisker in the fully charged graphitic foam clearly revealed Al and Cl Auger signals uniformly distributed over the whisker (Fig. 4d, e), again confirming chloroaluminate anion intercalation. When fully discharged, AES mapping revealed anion de-intercalation from graphite with much lower Al and Cl Auger signals observed (Fig. 4f, g). These spectroscopic results clearly revealed chloroaluminate ion intercalation/de-intercalation in the graphite redox reactions involved in our rechargeable Al cell.

The Al battery pouch cell is mechanically bendable and foldable (Supplementary Video 1) owing to the flexibility of the electrode and separator materials. Further, we drilled through Al battery pouch cells during battery operation and observed no safety hazard, owing to the lack of flammability of the ionic liquid electrolyte in air (see Supplementary Video 2).

We have developed a new Al-ion battery using novel graphitic cathode materials with a stable cycling life up to 7,500 charge/discharge cycles without decay at ultrahigh current densities. The present Al/graphite battery can afford an energy density of $\sim$40 W h kg^{-1} (comparable to lead–acid and Ni–MH batteries, with room for improvement by optimizing the graphitic electrodes and by developing other novel cathode materials) and a high power density, up to 3,000 W kg^{-1} (similar to supercapacitors). We note that the energy/power densities were calculated on the basis of the measured $\sim$65 mA h g^{-1} cathode capacity and the mass of active materials in electrodes and electrolyte. Such rechargeable Al ion batteries have the potential to be cost effective and safe, and to have high power density.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Yang, Z. *et al.* Electrochemical energy storage for green grid. *Chem. Rev.* **111,** 3577–3613 (2011).
2. Huskinson, B. *et al.* A metal-free organic–inorganic aqueous flow battery. *Nature* **505,** 195–198 (2014).
3. Li, Q. & Bjerrum, N. J. Aluminum as anode for energy storage and conversion: a review. *J. Power Sources* **110,** 1–10 (2002).
4. Gifford, P. R. & Palmisano, J. B. An aluminum/chlorine rechargeable cell employing a room temperature molten salt electrolyte. *J. Electrochem. Soc.* **135,** 650–654 (1988).
5. Jayaprakash, N., Das, S. K. & Archer, L. A. The rechargeable aluminum-ion battery. *Chem. Commun.* **47,** 12610–12612 (2011).
6. Rani, J. V., Kanakaiah, V., Dadmal, T., Rao, M. S. & Bhavanarushi, S. Fluorinated natural graphite cathode for rechargeable ionic liquid based aluminum-ion battery. *J. Electrochem. Soc.* **160,** A1781–A1784 (2013).
7. Hudak, N. S. Chloroaluminate-doped conducting polymers as positive electrodes in rechargeable aluminum batteries. *J. Phys. Chem. C* **118,** 5203–5215 (2014).
8. Armand, M. & Tarascon, J. M. Building better batteries. *Nature* **451,** 652–657 (2008).
9. Yu, X., Lu, B. & Xu, Z. Super long-life supercapacitors based on the construction of nanohoneycomb-like strongly coupled CoMoO₄–3D graphene hybrid electrodes. *Adv. Mater.* **26,** 1044–1051 (2014).
10. Chen, Z. *et al.* Three-dimensional flexible and conductive interconnected graphene networks grown by chemical vapour deposition. *Nature Mater.* **10,** 424–428 (2011).
11. Wasserscheid, P. & Keim, W. Ionic liquids—new ''solutions'' for transition metal catalysis. *Angew. Chem. Int. Edn* **39,** 3772–3789 (2000).
12. Auborn, J. J. & Barberio, Y. L. An ambient temperature secondary aluminum electrode: its cycling rates and its cycling efficiencies. *J. Electrochem. Soc.* **132,** 598–601 (1985).
13. Wilkes, J. S., Levisky, J. A., Wilson, R. A. & Hussey, C. L. Dialkylimidazolium chloroaluminate melts: a new class of room-temperature ionic liquids for electrochemistry, spectroscopy and synthesis. *Inorg. Chem.* **21,** 1263–1264 (1982).
14. Lai, P. K. & Skyllas-Kazacos, M. Electrodeposition of aluminium in aluminium chloride/1-methyl-3-ethylimidazolium chloride. *J. Electroanal. Chem. Interfacial Electrochem.* **248,** 431–440 (1988).
15. Jiang, T., Chollier Brym, M. J., Dubé, G., Lasia, A. & Brisard, G. M. Electrodeposition of aluminium from ionic liquids: Part I—electrodeposition and surface morphology of aluminium from aluminium chloride (AlCl₃)–1-ethyl-3-methylimidazolium chloride ([EMIm]Cl) ionic liquids. *Surf. Coat. Tech.* **201,** 1–9 (2006).
16. Borg, R. J. & Dienes, G. J. *An Introduction to Solid State Diffusion* (Academic, 1988).
17. Zhu, Y.-J., Hansen, T. A., Ammermann, S., McBride, J. D. & Beebe, T. P. Nanometer-size monolayer and multilayer molecule corrals on HOPG: a depth-resolved mechanistic study by STM. *J. Phys. Chem. B* **105,** 7632–7638 (2001).
18. Schmuelling, G. *et al.* X-ray diffraction studies of the electrochemical intercalation of bis(trifluoromethanesulfonyl)imide anions into graphite for dual-ion cells. *J. Power Sources* **239,** 563–571 (2013).
19. Takahashi, S., Koura, N., Kohara, S., Saboungi, M. L. & Curtiss, L. A. Technological and scientific issues of room-temperature molten salts. *Plasmas Ions* **2,** 91–105 (1999).
20. Hardwick, L. J. *et al.* An in situ Raman study of the intercalation of supercapacitor-type electrolyte into microcrystalline graphite. *Electrochim. Acta* **52,** 675–680 (2006).

METHODS

Preparation of ionic liquid electrolytes. A room temperature ionic liquid electrolyte was made by mixing 1-ethyl-3-methylimidazolium chloride ([EMIm]Cl, 97%, Acros Chemicals) and anhydrous aluminium chloride (AlCl₃, 99.999%, Sigma Aldrich). [EMIm]Cl was baked at 130 °C under vacuum for 16–32 h to remove residual water. ([EMIm]AlₓClᵧ) ionic liquid electrolytes were prepared in an argon-atmosphere glove box (both [EMIm]Cl and AlCl₃ are highly hygroscopic) by mixing anhydrous AlCl₃ with [EMIm]Cl, and the resulting light-yellow, transparent liquid was stirred at room temperature for 10 min. The mole ratio of AlCl₃ to [EMIm]Cl was varied from 1.1 to 1.8. The water content of the ionic liquid was determined (500–700 p.p.m.) using a coulometric Karl Fischer titrator, DL 39 (Mettler Toledo). The predominant anions in basic melts (AlCl₃/[EMIm]Cl mole ratio <1) are Cl⁻ and AlCl₄⁻, while in acidic melts (AlCl₃/[EMIm]Cl mole ratio >1) chloroaluminate anions such as Al₂Cl₇⁻, Al₃Cl₁₀⁻, and Al₄Cl₁₃⁻ are formed[11]. The ratio of anions to cations in the AlCl₃/[EMIm]Cl electrolyte was determined using a glass fibre filter paper (Whatman GF/D) loaded with a 4–8 μm Au-coated SiO₂ beads[21] in a cuvette cell (0.35 ml, Starna Cells) with random orientation quartz windows. Then, in the glove box, the cuvette cell was filled with AlCl₃/[EMIm]Cl = 1.3 (by mole). Raman spectra (200–650 cm⁻¹) were obtained using a 785-nm laser with 2 cm⁻¹ resolution. Raman data were collected from the surface of the Au-coated SiO₂ bead so as to benefit from surface enhanced Raman[21,22] (Extended Data Fig. 2b).

Preparation of graphitic foam. Nickel (Ni) foams (Alantum Advanced Technology Materials, Shenyang, China), were used as 3D scaffold templates for the CVD growth of graphitic foam, following the process reported previously[9,10]. The Ni foams were heated to 1,000 °C in a horizontal tube furnace (Lindberg Blue M, TF55030C) under Ar (500 standard cubic centimetres per minute or s.c.c.m.) and H₂ (200 s.c.c.m.) and annealed for 10 min to clean their surfaces and to eliminate a thin surface oxide layer. Then, methane (CH₄) was introduced into the reaction tube at ambient pressure at a flow rate of 10 s.c.c.m., corresponding to a concentration of 1.4 vol.% in the total gas flow. After 10 min of reaction gas mixture flow, the samples were rapidly cooled to room temperature at a rate of 300 °C min⁻¹ under Ar (500 s.c.c.m.) and H₂ (200 s.c.c.m.). The Ni foams covered with graphite were drop-coated with a poly(methyl methacrylate) (PMMA) solution (4.5% in ethyl acetate), and then baked at 110 °C for 0.5 h. The PMMA/graphene/Ni foam structure was obtained after solidification. Afterwards, these samples were put into a 3 M HCl solution for 3 h to completely dissolve the Ni foam to obtain the PMMA/graphite at 80 °C. Finally, the pure graphitic foam was obtained by removing PMMA in hot acetone at 55 °C and annealing in NH₃ (80 s.c.c.m.) at 600 °C for 2 h, and then annealing in air at 450 °C for 2 h. The microstructure of the graphitic foam was examined by SEM analysis using a FEI XL30 Sirion scanning electron microscope (Fig. 2a in the main text).

Preparation of glassy carbon. Glassy carbon (GC) was used as the current collector in the Swagelok-type cell. 72 g phenol (Sigma-Aldrich) and 4.5 ml ammonium hydroxide (30%, Fisher Scientific) were dissolved in 100 ml formaldehyde solution (37%, Fisher Scientific) under reflux while stirring rapidly. The solution was stirred at 90 °C until the solution turned a milk-white colour. Rotary evaporation was used to remove the water and get the phenolic resin. The phenolic resin was solidified at 100 °C in a mould (1/2-inch glass tube), and then carbonized at 850 °C under an Ar atmosphere for four hours to obtain the GC rod. The resulting GC rod contributed negligible capacity to the cathode (Extended Data Fig. 6b).

Electrochemical measurements. Prior to assembling the Al/graphite cell in the glove box, all components were heated under vacuum at 60 °C for more than 12 h to remove residual water. All electrochemical tests were performed at 25 ± 1 °C. A Swagelok-type cell (1/2 inch diameter) was constructed using a ~4 mg PG foil (0.017 mm, Suzhou Dasen Electronics Materials) cathode and a 90 mg Al foil (0.25 mm, Alfa Aesar) anode. A 1/2 inch GC rod (10 mm) was used as the current collector for the PG cathode, and a 1/2 inch graphite rod (10 mm) was used for the Al anode. Six layers of 1/2 inch glass fibre filter paper (Whatman 934-AH) were placed between the anode and cathode. Then, ~1.0 ml of ionic liquid electrolyte (prepared with AlCl₃/[EMIm]Cl mole ratios of 1.1, 1.3, 1.5 and 1.8) was injected and the cell sealed. The Al/PG cell was then charged (to 2.45 V) and discharged (to 0.01 V) at a current density of 66 mA g⁻¹ with a MTI battery analyser (BST8-WA) to identify the ideal AlCl₃/[EMIm]Cl mole ratio (Extended Data Fig. 2a). To investigate the Coulombic efficiency of the Al/PG cell in AlCl₃/[EMIm]Cl ≈ 1.3 (by mole) electrolyte, the cell was charged to 2.45, 2.50, 2.55 and 2.60 V, respectively, and discharged to 0.4 V at a current density of 66 mA g⁻¹ (Extended Data Fig. 6a). For long-term cycling stability tests, an Al/PG cell using electrolyte AlCl₃/[EMIm]Cl ≈ 1.3 by mole was charged/discharged at a current density of 66 mA g⁻¹ (Fig. 1b, c in the main text). To study the rate capability of the Al/PG cell, the current densities were varied from 66 to 264 mA g⁻¹ (Extended Data Fig. 7). Note that we lowered the electrolyte amount to ~0.02 ml per mg of cathode material and observed similar cell operation (Extended Data Fig. 4). Further decrease in the electrolyte ratio is possible through battery engineering.

PG foil was synthesized by pyrolysis of polyimide at high temperature, in which some covalent bonding is inevitably generated due to imperfections. Natural graphite foil was produced by compressing expanded graphite flakes, leading to stacking of natural graphite flakes by Van der Waals bonding between them. Similar battery characteristics were observed with PG and graphite foil electrodes, indicating that the battery behaviour was derived from the graphitic property of the electrodes (Extended Data Fig. 8c). However, since the natural graphite foils are synthesized by compressing expanded natural graphite powders without the covalent linkage between them, these foils suffered from drastic electrode expansion obvious to the naked eye, whereas pyrolytic graphite foils showed no obvious electrode expansion due to covalency (Extended Data Fig. 8a, b).

Pouch cells were assembled in the glove box using a graphitic-foam (~3 mg) cathode and an Al foil (~70 mg) anode, which were separated by two layers of glass fibre filter paper to prevent shorting. Polymer (0.1 mm × 4 mm × 5 mm) coated Ni foils (0.09 mm × 3 mm × 60 mm in size; MTI corporation) were used as current collectors for both anode and cathode. The electrolyte (~2 ml prepared using AlCl₃/[EMIm]Cl = 1.3 by mole) was injected and the cell was closed using a heat sealer. The cell was removed from the glove box for long-term cycling stability tests, in which the cell was charged/discharged at a current density of 4,000 mA g⁻¹ (Fig. 2b, c). To determine the rate capability and fast-charge/slow-discharge behaviours of the Al/graphitic-foam cell, various current densities from 100 to 5,000 mA g⁻¹ were used (Extended Data Fig. 9 and Fig. 2d). The pouch cell was charged to 2.42 V and discharged to a cut-off voltage of 0.5 V to prevent the dissolution reaction of Ni foil in the ionic liquid electrolyte.

Cyclic voltammetry measurements were performed using a potentiostat/galvanostat model CHI 760D (CH Instruments) in either three-electrode or two-electrode mode. The working electrode was an Al foil or a PG foil, the auxiliary electrode consisted of an Al foil, and an Al foil was used as the reference electrode. Copper tape (3M) was attached to these electrodes as the current collector. The copper tape was covered by poly-tetrafluoroethylene (PTFE) tape to prevent contact with the ionic liquid electrolyte and the part of the copper tape covered by PTFE was not immersed in the ionic liquid electrolyte. This prevented corrosion of the copper tape during cyclic voltammetry measurements. All three electrodes were placed in a plastic (1.5 ml) cuvette cell (containing electrolyte AlCl₃/[EMIm]Cl= 1.3 by mole) in the glove box, and then sealed with a rubber cap using a clamp. The scanning voltage range was set from –1.0 to 1.0 V (versus Al) for Al foil and 0 to 2.5 V (versus Al) for graphitic material, and the scan rate was 10 mV s⁻¹ (Extended Data Fig. 10d). To investigate the working voltage range of the electrolyte without involving cathode intercalation, two-electrode measurement was performed by using a GC rod cathode against an Al anode in a Swagelok cell in AlCl₃/[EMIm]Cl (~1.3 by mole) electrolyte. The scanning voltage range was set from 0 to 2.9 V at a scan rate of 10 mV s⁻¹ (Extended Data Fig. 6b).

We investigated the Al ion cell operation mechanism and electrode reactions in the ionic liquid electrolyte, using the optimal mole ratio of AlCl₃/[EMIm]Cl = 1.3. Using CV (Extended Data Fig. 10d), a reduction wave from –1.0 to –0.08 V (versus Al) and an oxidation wave from −0.08 to 0.80 V (versus Al) for the anode were observed (Extended Data Fig. 10d, left plot), corresponding to Al reduction/electrodeposition and oxidation/dissolution[13,15,23–25] during charging and discharging, respectively. This was consistent with Al redox electrochemistry in chloroaluminate ionic liquids[13,15,23–25] via equation (1) in the main text, and consistent with our Raman measurements, which showed both AlCl₄⁻ and Al₂Cl₇⁻ in the electrolyte (Extended Data Fig. 2b). On the graphitic cathode side, an oxidation wave of 1.83 to 2.50 V (versus Al) and a reduction wave of 1.16 to 2.36 V (versus Al) were observed (Extended Data Fig. 10d, right plot) and attributed to graphite oxidation and reduction through intercalation and de-intercalation of anions (predominantly AlCl₄⁻ due to its smaller size), respectively. The oxidation voltage range of 1.83 to 2.50 V (versus Al, Extended Data Fig. 10d, right plot) was close to the anodic voltage range (1.8 to 2.2 V versus Al) of a previously reported dual-graphite cell[26] attributed to AlCl₄⁻ intercalation in graphite. The reduction wave range of 1.16 to 2.36 V (versus Al) was assigned to the AlCl₄⁻ de-intercalation[26]. The nature of the shoulder in the reduction curve of graphite ranging from 2.36 to 1.9 V (Extended Data Fig. 10d, right plot) and a higher discharge plateau (2.25 to 2.0 V) of an Al/PG cell upon charging (Fig. 1b in the main text) remained unclear, but could be due to different stages of anion–graphite intercalation[27].

XRD and Raman studies of graphite cathodes during charge and discharge. For *ex situ* X-ray diffraction (XRD) study, an Al/ PG cell (in a Swagelok configuration) was charged and discharged at a constant current density of 66 mA g⁻¹. The reactions were stopped after 30 mA h g⁻¹ charged, fully charged (62 mA h g⁻¹) and 40 mA h g⁻¹ discharged after charge/discharge capacities were in a stable state. Fully charged (62 mA h g⁻¹) graphitic foam was also prepared. After either the charge or the discharge reaction, the graphitic cathode was removed from the cell in the glove box. To avoid reaction between the cathode and air/moisture in the ambient atmosphere, the cathode was placed onto a glass slide and then wrapped in a Scotch tape.

The wrapped samples were immediately removed from the glove box for *ex situ* XRD measurements, which were performed on a PANalytical X'Pert instrument (Fig. 3a in the main text and Extended Data Fig. 10b).

The periodic repeat distance (I_C), the intercalant gallery height (d_i) and the gallery expansion (Δd)[28,29] can be calculated using

$$I_C = (d_i + 3.35\,\text{Å}) \times (n-1) = (\Delta d + 3.35\,\text{Å}) \times n = l \times d_{obs} \quad (3)$$

where l is the index of $(00l)$ planes oriented in the stacking direction and d_{obs} is the observed value of the spacing between two adjacent planes[18,28,29]. The d spacing of graphite is 3.35 Å. During the charging/anion-intercalation process, the graphite (002) peak completely vanished and two new peaks arose. The intensity pattern is commonly found for a stage n graphite intercalation compound (GIC), where the most dominant peak is the $(00n + 1)$ and the second most dominant peak is the $(00n + 2)$[18,28,29]. Based on our experimental data, by increasing the charging state from 48–60% charged (30 mA h g^{-1}) to the fully charged state (62 mA h g^{-1}), the distance between the $(00n + 1)$ and $(00n + 2)$ peaks gradually increased, as more $Al_xCl_y^-$ anions intercalated. The d spacing values of $(00n + 1)$ and $(00n + 2)$ peaks (that is, $d_{(n+1)}$ and $d_{(n+2)}$, respectively) were calculated from XRD data (for example, Fig. 3a). By determining the ratio of the $d_{(n+2)}/d_{(n+1)}$ peak position and correlating these to the ratios of stage pure GICs (that is, ideal cases), the most dominant stage phase of the observed GIC can be assigned[28,29]. After assigning the $(00l)$ indices, we calculated the intercalant gallery height (d_i) through equation (3).

For simultaneous *in situ* Raman and galvanostatic charge/discharge reaction measurements, a cuvette cell (0.35 ml, Starna Cells) with random orientation quartz windows was used. An aluminium foil and a graphitic material (PG or graphitic foam) were used as the anode and cathode, respectively. The electrolyte was mixed AlCl$_3$/[EMIm]Cl = 1.3 (by mole). The electrochemical cell was assembled in the glove box following the process mentioned above. Raman spectra were obtained (1,500–1,700 cm^{-1}) using a HeNe laser (633 nm) with 2 cm^{-1} resolution. The spectral data were collected after a few successive charge/discharge scans between 2.45 and 0.01 V at a current density of 66 mA g^{-1} (PG) (Fig. 3b in the main text) or 1,000 mA g^{-1} (graphitic foam) (Extended Data Fig. 10a).

XPS and AES measurements. Al/graphitic-foam cells were fully charged/discharged at a current density of 4,000 mA g^{-1}. Then, the Al/graphitic-foam cells were transferred to the glove box for preparation for XPS and AES analysis. Fully charged/discharged graphitic foams were collected from the pouch cell and washed with anhydrous methanol to remove the residual AlCl$_3$/EMIC ionic liquid electrolyte. The as-rinsed graphitic foams were attached to a Si wafer and baked at 90 °C for 10 min to remove residual methanol. The samples were sealed in a plastic pouch to avoid contamination by reaction with moisture and oxygen before XPS and AES characterization. Auger electron spectra were taken by a PHI 700 Scanning Auger Nanoprobe operating at 10 kV and 10 nA. XPS spectra were collected on a PHI VersaProbe Scanning XPS Microprobe (Fig. 4 in the main text).

TGA measurements. Fully charged PG cathodes were washed with methanol for 24 h to remove the residual AlCl$_3$/EMIC ionic liquid electrolyte. The as-washed PG samples were calcined at 850 °C for 3 h in air. The as-calcined samples (white foam) were collected, weighed, and analysed by SEM-EDX to study the chemical composition (Extended Data Fig. 10c). SEM and SEM-EDX analyses were performed using an FEI XL30 Sirion scanning electron microscope.

Sample size. No statistical methods were used to predetermine sample size.

21. Zhang, B. *et al.* Plasmonic micro-beads for fluorescence enhanced, multiplexed protein detection with flow cytometry. *Chem. Sci.* **5,** 4070–4075 (2014).
22. Tabakman, S. M., Chen, Z., Casalongue, H. S., Wang, H. & Dai, H. A new approach to solution-phase gold seeding for SERS substrates. *Small* **7,** 499–505 (2011).
23. Lee, J. J., Bae, I. T., Scherson, D. A., Miller, B. & Wheeler, K. A. Underpotential deposition of aluminum and alloy formation on polycrystalline gold electrodes from AlCl$_3$/EMIC room-temperature molten salts. *J. Electrochem. Soc.* **147,** 562–566 (2000).
24. Pan, S.-J., Tsai, W.-T., Chang, J.-K. & Sun, I. W. Co-deposition of Al–Zn on AZ91D magnesium alloy in AlCl$_3$–1-ethyl-3-methylimidazolium chloride ionic liquid. *Electrochim. Acta* **55,** 2158–2162 (2010).
25. Endres, F., MacFarlane, D. & Abbott, A. *Electrodeposition from Ionic Liquids* (Wiley & Sons, 2008).
26. Carlin, R. T., De Long, H. C., Fuller, J. & Trulove, P. C. Dual intercalating molten electrolyte batteries. *J. Electrochem. Soc.* **141,** L73–L76 (1994).
27. Bao, W. *et al.* Approaching the limits of transparency and conductivity in graphitic materials through lithium intercalation. *Nature Commun.* **5,** 4224 (2014).
28. Zhang, X., Sukpirom, N. & Lerner, M. M. Graphite intercalation of bis(trifluoromethanesulfonyl) imide and other anions with perfluoroalkanesulfonyl substituents. *Mater. Res. Bull.* **34,** 363–372 (1999).
29. Özmen-Monkul, B. & Lerner, M. M. The first graphite intercalation compounds containing tris(pentafluoroethyl)trifluorophosphate. *Carbon* **48,** 3205–3210 (2010).

LETTER

Multistep continuous–flow synthesis of (R)– and (S)-rolipram using heterogeneous catalysts

Tetsu Tsubogo[1], Hidekazu Oyamada[1] & Shū Kobayashi[1]

Chemical manufacturing is conducted using either batch systems or continuous-flow systems. Flow systems have several advantages over batch systems, particularly in terms of productivity, heat and mixing efficiency, safety, and reproducibility[1–4]. However, for over half a century, pharmaceutical manufacturing has used batch systems because the synthesis of complex molecules such as drugs has been difficult to achieve with continuous-flow systems[5,6]. Here we describe the continuous-flow synthesis of drugs using only columns packed with heterogeneous catalysts. Commercially available starting materials were successively passed through four columns containing achiral and chiral heterogeneous catalysts to produce (R)-rolipram[7], an anti-inflammatory drug and one of the family of γ-aminobutyric acid (GABA) derivatives[8]. In addition, simply by replacing a column packed with a chiral heterogeneous catalyst with another column packed with the opposing enantiomer, we obtained antipole (S)-rolipram. Similarly, we also synthesized (R)-phenibut, another drug belonging to the GABA family. These flow systems are simple and stable with no leaching of metal catalysts. Our results demonstrate that multistep (eight steps in this case) chemical transformations for drug synthesis can proceed smoothly under flow conditions using only heterogeneous catalysts, without the isolation of any intermediates and without the separation of any catalysts, co-products, by-products, and excess reagents. We anticipate that such syntheses will be useful in pharmaceutical manufacturing.

Although the chemical and biotechnology industries have preferred to use continuous-flow systems because of their high productivity and efficiency, fine chemical production has been conducted using batch systems because the synthesis of more-complex molecules has been difficult to achieve with continuous-flow systems. However, recently pharmaceutical manufacturing has begun to require high quality of synthesis, environmentally benign methods and reproducibility of manufacturing. To meet these demands, it is believed that continuous-flow systems are superior to batch systems.
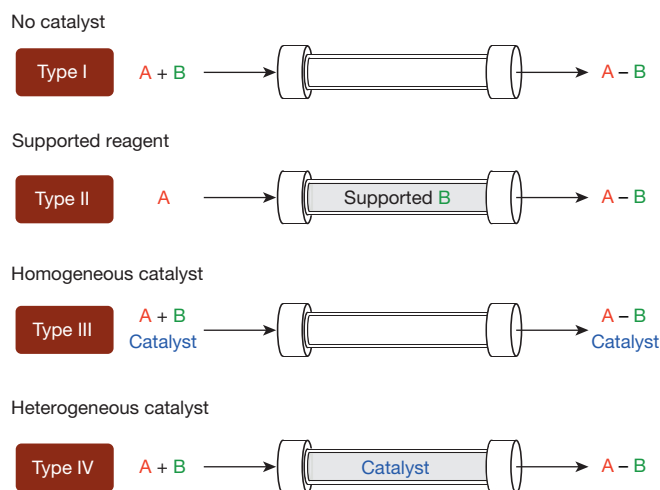
Methods for continuous-flow systems have been developed more recently than methods for batch systems. We divided the continuous-flow systems into four types (I–IV; see Fig. 1). In type I, substrates (A and B) are passed through a column or hollow loop, inside which reactions occur. Unreacted A or B or any by-products are not separated. In type II, one of the substrates (B) is supported in a column. If an excess amount of B is used, one substrate (A) is consumed. However, once the supported B is consumed, the column must be changed. In type III, A reacts with B in the presence of a homogeneous catalyst. Although catalysis proceeds smoothly, the catalyst cannot be separated. In type IV, A reacts with B in the presence of a heterogeneous catalyst. If catalysis proceeds smoothly, no separation is required.

The recent regulations of 'green sustainable chemistry'[9] mean that synthesis with catalysts is preferable to synthesis without catalysts because of energy savings and waste reduction. Consequently, types III and IV are recommended in continuous-flow systems. Furthermore, although catalysts are contaminated with products in type III, no contamination of catalysts is expected under ideal conditions in

type IV. Therefore, given that type IV is regarded as the best method for continuous-flow synthesis[10–13], we elected to use type IV for our drug synthesis. Although recent technological improvements have made it possible to synthesize relatively complex molecules, including drugs, using continuous-flow systems[14–17], there have been no examples of drug synthesis using only type IV continuous-flow systems.

γ-Aminobutyric acid (GABA) and its derivatives are an important class of compounds in neuroscience[8]. Rolipram is one of the GABA family. It is an anti-inflammatory drug[7,18]—a selective phosphodiesterase 4 (PDE4) inhibitor and particularly effective for the PDE4B subtype of PDE4[19]. Moreover, rolipram is known to be a possible antidepressant and has been reported to have anti-inflammatory, immunosuppressive, and antitumour effects[7,18]. Rolipram has also been proposed as a treatment for multiple sclerosis, and has been suggested to have antipsychotic effects[20]. Furthermore, it has been reported that (R)-rolipram has anti-inflammatory activity, whereas (S)-rolipram does not[19]. There are many GABA derivatives that are drugs or have potential biological activities in the area of neurotransmitters and brain science[21] (Fig. 2).

We selected (R)- and (S)-rolipram for the target of our continuous-flow synthesis because rolipram itself is a very promising drug in several ways and because the completed flow synthesis may be applicable to the synthesis of other GABA derivatives. We planned to synthesize (R)- and (S)-rolipram from commercially available starting materials using continuous-flow systems, using only type IV columns (Fig. 1). Our synthetic strategy is shown in Fig. 3. Commercially available aldehyde **2** and nitromethane **3** could be converted to nitroalkene **4**. Catalytic asymmetric 1,4-addition of malonate **5** to **4** could afford

Figure 1 | The four types of continuous-flow systems. The continuous-flow systems so far reported can be divided into types I–IV, as illustrated, using substrates A and B. See main text for details. Type IV is regarded as the best method for continuous-flow synthesis.

---

[1]Department of Chemistry and Green & Sustainable Chemistry Social Cooperation Laboratory, School of Science, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.
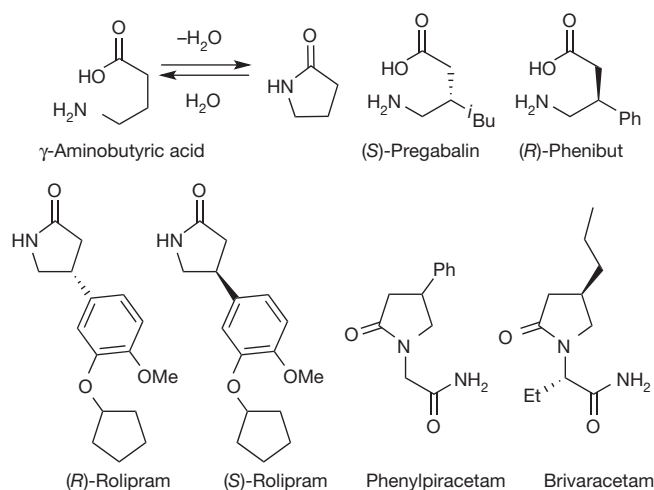
Figure 2 | GABA derivatives. These drugs and potential drugs are described in the main text.

enantiomerically enriched γ-nitro ester **6**. The nitro group of **6** could be reduced selectively to afford γ-lactam **7** after cyclization. Finally, the ester group of **7** could be removed to afford **1**.

First, we examined the flow synthesis of **4** from **2** and **3**, using a heterogeneous catalyst (Fig. 4, stage 1)[22,23]. The formation of nitroalkenes from aldehydes and nitroalkanes is known to proceed in the presence of a base[24]. We selected toluene as a solvent because the following step, the asymmetric 1,4-addition, proceeded smoothly in toluene. We examined several heterogeneous amines, and finally found that a silica-supported amine with anhydrous calcium chloride showed a high yield of **4** when using almost equimolar amounts of **2** and **3** at 50 °C–75 °C. Under these optimized conditions, a silica-supported amine (Chromatorex DM1020; Fuji Silysia; 4.5 g, 0.73 mmol g⁻¹) and finely crushed anhydrous calcium chloride (13.5 g) were introduced into a SUS (stainless steel) column (diameter 10 mm, length 300 mm; column I). The toluene solution of **2** and **3** was introduced from the bottom of the column, and the desired product **4** was obtained in >90% yield. The system was found to be stable at 75 °C for at least one week (>90% yield). We further confirmed that
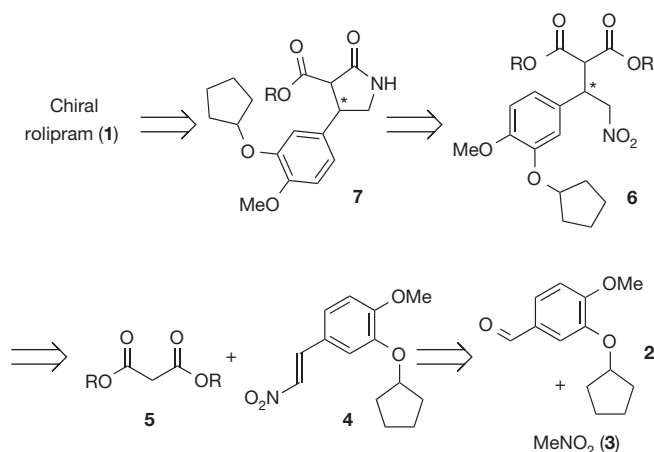

Figure 3 | Retrosynthetic analysis. Commercially available aldehyde **2** could react with nitromethane **3** to afford nitroalkene **4**. Asymmetric 1,4-addition of malonate **5** to **4** with a chiral catalyst could give enantiomerically enriched γ-nitro ester **6**, whose nitro group could be selectively reduced to afford γ-lactam **7** after cyclization. Finally, the ester group of **7** could be removed to afford optically active rolipram **1**. (*R*)- and (*S*)-rolipram could be synthesized by using column II packed with PS-(*S*)-pybox-calcium chloride and column II′ bearing PS-(*R*)-pybox-calcium chloride (the opposing enantiomer), respectively.

this flow system was applicable to other aldehydes; several nitroalkenes were obtained in high yields (Supplementary Information). We noted that, although excess amounts of aldehydes (1.5–1.7 equivalents) were required to obtain high yields of nitroalkenes in batch systems, almost equimolar amounts of aldehydes afforded the desired nitroalkenes in high yields under continuous-flow conditions.

We next examined the asymmetric 1,4-addition of malonate **5** to **4** using a chiral heterogeneous catalyst (Fig. 4, stage 2). Catalytic asymmetric reactions provide one of the most efficient routes to enantiomerically enriched products[25]. Recently, we developed a polymer-supported chiral calcium catalyst, which was successfully used for the asymmetric 1,4-addition of malonates to nitroalkenes under continuous-flow conditions[26]. We set column II, which was filled with this polymer-supported (PS) calcium catalyst (PS-(*S*)-pybox-calcium chloride, where pybox is pyridinebisoxazoline), and connected it with column I. We also included a valve (switching paths A and B) to drain the synthesized nitroalkene solution (receiver 1) and an MS 4A column (column X, diameter 5 mm, length 50 mm) to stabilize the system. A solution of nitroalkene **4** synthesized in column I and a toluene solution of malonate **5** and triethylamine were mixed and introduced into column II. After optimization of the reaction conditions, it was found that when the reaction was conducted at 0 °C, using slightly excess amounts of nitromethane **3** and malonate **5** (**4** was formed), the desired γ-nitro ester **6** was obtained in high yield with high enantioselectivity. Under the optimized conditions, the mixture of **2**, **3**, and **5** was precooled at 0 °C using a loop, and column II was separated into two columns (column II-1 and column II-2, each of diameter 10 mm and length 100 mm, packed with 750 mg of 0.85 mmol g⁻¹ PS-(*S*)-pybox, 375 mg CaCl₂•2H₂O and 1.4 g Celite; this division of the column was required due to the size of the cooling bath). We collected the crude product solution in receiver 2. It was confirmed to contain mainly **6**, with small amounts of **3** and **5**, and triethylamine. The crude product was quenched with a solid ammonium chloride, and after a usual work-up, the desired γ-nitro ester **6** was obtained in 84% yield with 94% enantiomeric excess. At this stage, we also tested several aldehydes in this continuous-flow system. It was found that, in all cases, the desired γ-nitro esters were obtained in high yields with high enantioselectivities (Supplementary Information).

The next step involved the reduction of the nitro group to the corresponding amino group (Fig. 4, stage 3). Experimental conditions required the flow of the toluene solution obtained from column II to be under atmospheric pressure. We selected a continuous-flow hydrogenation[27,28] and examined several commercially available supported Ni and Pd catalysts[18,29]; however, the desired reduction did not proceed at all. Having recently developed a polysilane-supported palladium/alumina (Pd/PSi-Al₂O₃) catalyst, which worked well for the hydrogenation of alkenes, alkynes, and also nitrobenzene derivatives under flow conditions[30], we then tested Pd/PSi-Al₂O₃ for the hydrogenation of **6**. Unfortunately, the reaction did not proceed.

At this stage, therefore, we decided to develop a new heterogeneous catalyst for our purpose. After several trials, we developed a polysilane-supported palladium/carbon (Pd/DMPSi-C, where DMPSi is dimethylpolysilane) catalyst, which worked well for the reduction. We then connected column III (column III-1 and column III-2, both of diameter 10 mm and length 100 mm; packed with 4.8 g of 0.29 mmol g⁻¹ Pd/DMPSi-C and 1.2 g Celite) with the already constructed flow system (columns I and II). The mixed solution (crude **6** in toluene) and hydrogen gas (3 ml min⁻¹) were introduced into column III (filled with Pd/DMPSi-C and Celite) from the top, pumped downward at 100 °C. Under these conditions, the desired reduction proceeded smoothly to afford γ-lactam **7** in 74% yield with 94% enantiomeric excess. We note that the reduction of the nitro group proceeded smoothly under atmospheric pressure of hydrogen, and that no epimerization occurred under the conditions. We also tested other substrates and in all cases the reduction proceeded well to afford the
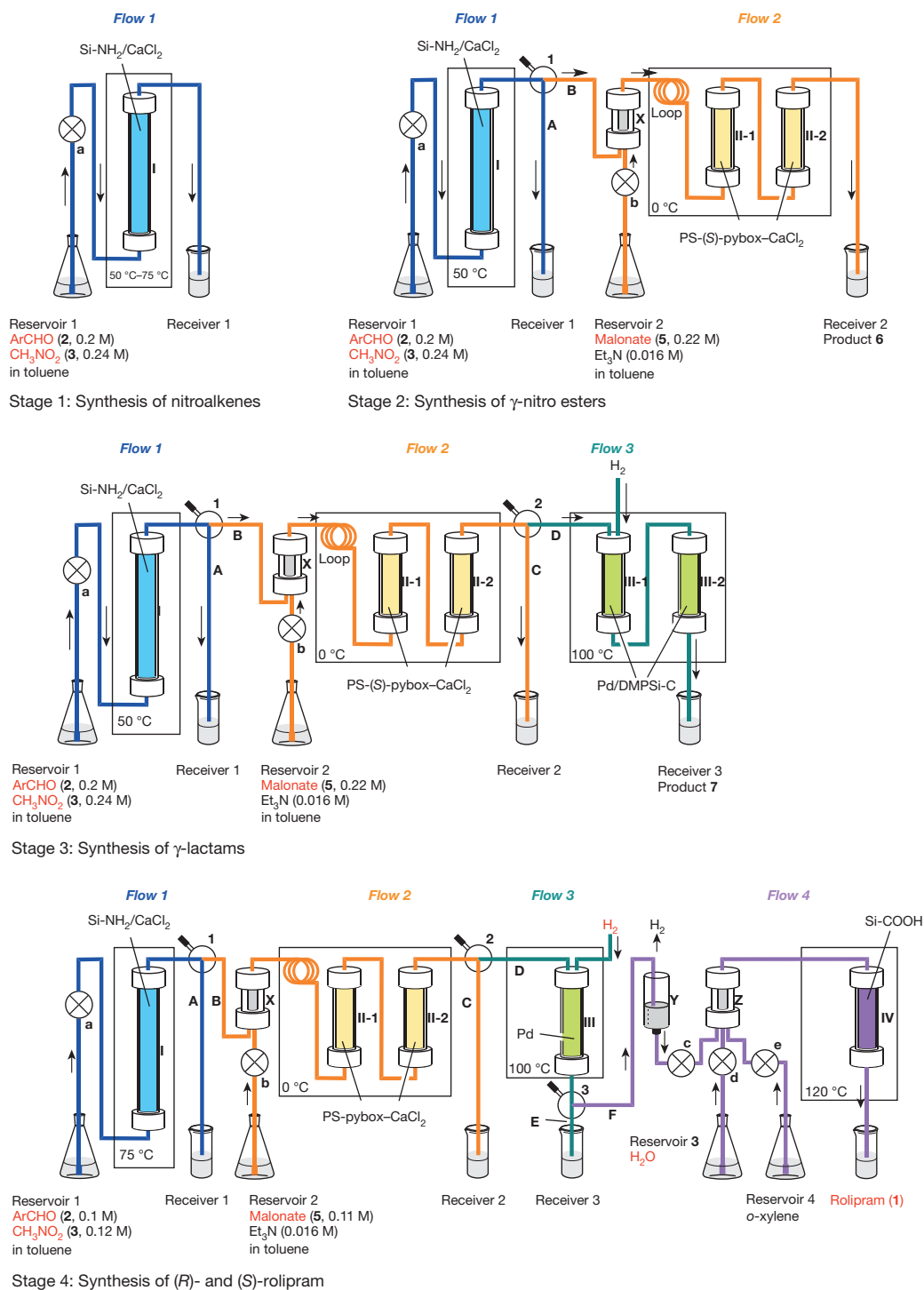
Figure 4 | Diagram of the series of flow reactors. Stage 1: Synthesis of nitroalkene **4** from aldehyde **2** and nitromethane **3**. Stage 2: Synthesis of γ-nitro ester **6** from aldehyde **2** and nitromethane **3**. (Et₃N is triethylamine.) Stage 3: Synthesis of γ-lactam **7** from aldehyde **2** and nitromethane **3**. Stage 4: Synthesis of (R)- and (S)-rolipram **1** from aldehyde **2** and nitromethane **3**. In total, the commercially available starting materials **2**, **3**, **5**, H₂, and H₂O were successively passed through columns I, II-1 and II-2, III, and IV containing heterogeneous achiral and chiral catalysts to directly afford **1** with high enantioselectivity.

Eight-step chemical transformations were conducted smoothly during the flow without isolation of any intermediates and without the separation of any catalysts, co-products, by-products, and excess reagents. We note that all four columns employed are the desirable type IV flow system (Fig. 1). Red text indicates starting materials and products. Structures labelled a, b, c, d and e are pumps; those labelled A, B, C, D, E and F are flow lines. X is MS 4A; Y is Amberlyst 15Dry; Z is Celite.

desired γ-lactams in high yields with high enantioselectivities (Supplementary Information).

The final stage in the synthesis of rolipram (**1**) involved the hydrolysis and decarboxylation of the ester part of **7** (Fig. 4, stage 4). We found that the desired transformations proceeded in the presence of a

silica-supported carboxylic acid (Si-COOH; Chromatorex ACD, Fuji Silysia). We then connected column IV (diameter 10 mm and length 300 mm), which was filled with Si-COOH (13.5 g, 0.38 mmol g⁻¹) and Celite (0.5 g), to columns I–III and examined the continuous flow starting from **2** and **3**. We then added small columns of Amberlyst

15Dry (column Y) and Celite (column Z), and *o*-xylene was introduced. The main flow from column III was combined with *o*-xylene and water, and the total flow was passed through column IV from the top down at 120 °C. Finally, we obtained (*S*)-rolipram ((*S*)-**1**, 50% yield from **2** after preparative thin layer chromatography, 997.8 mg per 24 h, 96% enantiomeric excess). The flow system was found to be stable for at least one week (Supplementary Information). Recrystallization from water/methyl alcohol gave optically pure (*S*)-rolipram (>99% enantiomeric excess). Direct recrystallization of the crude product afforded chemically and enantiomerically pure (*S*)-rolipram without chromatography.

Thus, the synthesis of (*S*)-rolipram was completed. Commercially available starting materials were successively passed through the columns containing heterogeneous achiral and chiral catalysts to produce the drug directly with high enantioselectivity. Eight-step chemical transformations were conducted smoothly during the flow without isolation of any intermediates and without the separation of any catalysts, co-products, by-products, and excess reagents. In the flow system, each step can be monitored by using receivers (real-time analysis is possible). We note that all four columns employed are the desirable type IV flow system (Fig. 1), and that the product does not contain any metal (palladium, <0.01 p.p.m.), as confirmed by inductively coupled plasma analysis. Moreover, this is the first example of the successful use of a chiral catalyst in multistep continuous-flow synthesis of drugs or biologically important compounds.

This flow system could also be applicable to the synthesis of other GABA derivatives (Fig. 2). Antipole (*R*)-rolipram was also synthesized by continuous flow by simply replacing column II packed with PS-(*S*)-pybox–calcium chloride with column II′ bearing PS-(*R*)-pybox–calcium chloride (the opposing enantiomer). The procedure remained the same and similar productivity was obtained ((*R*)-**1**, 50% yield from **2**, 96% enantiomeric excess). We also synthesized (*R*)-phenibut[21] from benzaldehyde by slightly modifying the flow system. We believe that all the compounds shown in Fig. 2 can be synthesized using continuous-flow systems.

The present multistep continuous-flow synthesis is at the laboratory scale, and the drugs were obtained on the gram scale. On the other hand, we have confirmed that the system is stable and the flow is at steady state during the synthesis. Indeed, the system is stable for at least one week, and the same yields and enantioselectivities were obtained for the syntheses of (*R*)- and (*S*)-rolipram. Furthermore, we also confirmed that heterogeneous catalysts used in this flow system are robust, air-stable, and have a long lifetime. For example, the chiral calcium catalyst can be used for several months or more without losing any catalytic activity and selectivity (enanitioselectivity). We are now scaling up the system towards multi-kilogram syntheses of drugs.

1. Wiles, C. & Watts, P. Continuous flow reactors: a perspective. *Green Chem.* **14**, 38–54 (2012).
2. Ley, S. V. & Baxendale, I. R. New tools and concepts for modern organic synthesis. *Nature Rev. Drug Discov.* **1**, 573–586 (2002).
3. Geyer, K., Codée, J. D. C. & Seeberger, P. H. Microreactors as tools for synthetic chemists—the chemists' round-bottomed flask of the 21st century? *Chemistry* **12**, 8434–8442 (2006).
4. Hartman, R. L., McMullen, J. P. & Jensen, K. F. Deciding whether to go with the flow: evaluating the merits of flow reactors for synthesis. *Angew. Chem. Int. Ed.* **50**, 7502–7519 (2011).
5. Van Arnum, P. Advancing flow chemistry in API manufacturing. *Pharm. Technol.* **37**, 78–82 (2013).
6. Poechlauer, P. *et al.* Continuous processing in the manufacture of active pharmaceutical ingredients and finished dosage forms: an industry perspective. *Org. Process Res. Dev.* **16**, 1586–1590 (2012).
7. Sommer, N. *et al.* The antidepressant rolipram suppresses cytokine production and prevents autoimmune encephalomyelitis. *Nature Med.* **1**, 244–248 (1995).
8. Macdonald, R. L. & Olsen, R. W. GABAA receptor channels. *Annu. Rev. Neurosci.* **17**, 569–602 (1994).
9. Anastas, P. T. & Warner, J. C. (eds) *Green Chemistry Theory and Practice* (Oxford Univ. Press, 1998).
10. Kirschning, A., Solodenko, W. & Mennecke, K. Combining enabling techniques in organic synthesis: continuous flow processes with heterogenized catalysts. *Chemistry* **12**, 5972–5990 (2006).
11. Frost, C. G. & Mutton, L. Heterogeneous catalytic synthesis using microreactor technology. *Green Chem.* **12**, 1687–1703 (2010).
12. Tsubogo, T., Ishiwata, T. & Kobayashi, S. Asymmetric carbon–carbon bond formation under continuous-flow conditions with chiral heterogeneous catalysts. *Angew. Chem. Int. Ed.* **52**, 6590–6604 (2013).
13. Battilocchio, C., Hawkins, J. M. & Ley, S. V. A mild and efficient flow procedure for the transfer hydrogenation of ketones and aldehydes using hydrous zirconia. *Org. Lett.* **15**, 2278–2281 (2013).
14. Pastre, J. C., Browne, D. L. & Ley, S. V. Flow chemistry syntheses of natural products. *Chem. Soc. Rev.* **42**, 8849–8869 (2013).
15. Hartwig, J. *et al.* Heating under high-frequency inductive conditions: application to the continuous synthesis of the neurolepticum olanzapine (Zyprexa). *Angew. Chem. Int. Ed.* **52**, 9813–9817 (2013).
16. Webb, D. & Jamison, T. F. Continuous flow multi-step organic synthesis. *Chem. Sci.* **1**, 675–680 (2010).
17. Hopkin, M. D., Baxendale, I. R. & Ley, S. V. An expeditious synthesis of imatinib and analogues utilising flow chemistry methods. *Org. Biomol. Chem.* **11**, 1822–1839 (2013).
18. Barnes, D. M. *et al.* Development of a catalytic enantioselective conjugate addition of 1,3-dicarbonyl compounds to nitroalkenes for the synthesis of endothelin-A antagonist ABT-546. Scope, mechanism, and further application to the synthesis of the antidepressant rolipram. *J. Am. Chem. Soc.* **124**, 13097–13105 (2002).
19. Day, J. P. *et al.* Elucidation of a structural basis for the inhibitor-driven, p62 (SQSTM1)-dependent intracellular redistribution of cAMP phosphodiesterase-4A4 (PDE4A4). *J. Med. Chem.* **54**, 3331–3347 (2011).
20. Maxwell, C. R., Kanes, S. J., Abel, T. & Siegel, S. J. Phosphodiesterase inhibitors: a novel mechanism for receptor-independent antipsychotic medications. *Neuroscience* **129**, 101–107 (2004).
21. Malykh, A. G. & Sadaie, M. R. Piracentam and piracetam-like drugs. *Drugs* **70**, 287–312 (2010).
22. Motokura, K., Tada, M. & Iwasawa, Y. Layered materials with coexisting acidic and basic sites for catalytic one-pot reaction sequences. *J. Am. Chem. Soc.* **131**, 7944–7945 (2009).
23. Soldi, L. *et al.* Use of immobilized organic base catalysts for continuous-flow fine chemical synthesis. *J. Catal.* **258**, 289–295 (2008).
24. Worrall, D. E. Nitrostyrene. *Org. Synth.* **9**, 66–69 (1929).
25. Ojima, I. (ed.) *Catalytic Asymmetric Synthesis* 3rd edn (Wiley, 2010).
26. Tsubogo, T., Yamashita, Y. & Kobayashi, S. Toward efficient asymmetric carbon–carbon bond formation: continuous flow with chiral heterogeneous catalysts. *Chemistry* **18**, 13624–13628 (2012).
27. Kobayashi, J. *et al.* A microfluidic device for conducting gas–liquid–solid hydrogenation reactions. *Science* **304**, 1305–1308 (2004).
28. O'Brien, M. *et al.* Hydrogenation in flow: homogeneous and heterogeneous catalysis using Teflon AF-2400 to effect gas-liquid contact at elevated pressure. *Chem. Sci.* **2**, 1250–1257 (2011).
29. Hynes, P. S., Stupple, P. A. & Dixon, D. J. Organocatalytic asymmetric total synthesis of (*R*)-rolipram and formal synthesis of (3*S*,4*R*)-paroxetine. *Org. Lett.* **10**, 1389–1391 (2008).
30. Oyamada, H., Naito, T. & Kobayashi, S. Continuous flow hydrogenation using polysilane-supported palladium/alumina hybrid catalysts. *Beilstein J. Org. Chem.* **7**, 735–739 (2011).

Author Contributions T.T. and H.O. designed and performed the experiments. S.K. conceived, designed and directed the investigations and wrote the manuscript with revisions provided by T.T.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.K. (shu_kobayashi@chem.s.u-tokyo.ac.jp).

Icebergs not the trigger for North Atlantic cold events

Stephen Barker[1], James Chen[1]†, Xun Gong[1], Lukas Jonkers[1], Gregor Knorr[2] & David Thornalley[3,4]

Abrupt climate change is a ubiquitous feature of the Late Pleistocene epoch[1]. In particular, the sequence of Dansgaard−Oeschger events (repeated transitions between warm interstadial and cold stadial conditions), as recorded by ice cores in Greenland[2], are thought to be linked to changes in the mode of overturning circulation in the Atlantic Ocean[3]. Moreover, the observed correspondence between North Atlantic cold events and increased iceberg calving and dispersal from ice sheets surrounding the North Atlantic[4] has inspired many ocean and climate modelling studies that make use of freshwater forcing scenarios to simulate abrupt change across the North Atlantic region and beyond[5−7]. On the other hand, previous studies[4,8] identified an apparent lag between North Atlantic cooling events and the appearance of ice-rafted debris over the last glacial cycle, leading to the hypothesis that iceberg discharge may be a consequence of stadial conditions rather than the cause[4,9−11]. Here we further establish this relationship and demonstrate a systematic delay between pronounced surface cooling and the arrival of ice-rafted debris at a site southwest of Iceland over the past four glacial cycles, implying that in general icebergs arrived too late to have triggered cooling. Instead we suggest that—on the basis of our comparisons of ice-rafted debris and polar planktonic foraminifera—abrupt transitions to stadial conditions should be considered as a nonlinear response to more gradual cooling across the North Atlantic. Although the freshwater derived from melting icebergs may provide a positive feedback for enhancing and or prolonging stadial conditions[10,11], it does not trigger northern stadial events.

We investigated fluctuations in surface ocean temperature and the delivery of ice-rafted debris (IRD) to a site in the northeast Atlantic (Ocean Drilling Program (ODP) site 983; 60.4° N, 23.6° W, 1,984 m depth; Fig. 1) at high temporal resolution (177 years on average) over the past ∼440 kyr (2,474 discrete samples). To this end we counted the relative proportion of the polar planktonic foraminifer, *Neogloboquadrina pachyderma*, within the total assemblage (%NPS; see Methods) and the number of lithogenic/terrigenous grains >150 μm per gram dry sediment (IRD per gram; see Methods). Today the location of ODP site 983 is under the influence of the warm surface Irminger Current (part of the modern subpolar gyre) as it turns northwards after splitting from the North Atlantic Current (NAC), which itself transports about 7.5 Sv (1 Sv = 10^6 m^3 s$^{−1}$) of warm (∼8.5 °C) water over the Iceland−Scotland Ridge and into the Nordic Seas[12] (Fig. 1). This inflow is balanced in part by the outflow of cold fresh surface waters via the East Greenland Current but predominantly (∼6 Sv) by overflows of cold dense bottom waters through the Denmark Strait and across the Iceland−Scotland Ridge that form as a result of strong wintertime cooling and convection within the Nordic Seas[12]. Together, these overflows represent the principal constituent precursors to North Atlantic Deep Water

(NADW) and therefore represent an essential component of the modern Atlantic Meridional Overturning Circulation (AMOC)[12].

The present ingress of warm NAC waters into the Nordic Seas is reflected by the southwest−northeast orientation of the North Atlantic polar front (Fig. 1). During the Last Glacial Maximum (LGM; ∼23,000−19,000 years ago, that is ∼23−19 kyr ago) the polar front was positioned much further south and was more zonally orientated[13] (Fig. 1), suggesting a reduction in heat transport into the Nordic Seas by the NAC. Palaeoceanographic reconstructions[14] and a range of model experiments[15] suggest that this difference was reflected by a change in the geometry of the AMOC, with the northern locus of deep water formation shifted to the south of Iceland. An analogous (though not identical) change is thought to have accompanied the abrupt shifts associated with stadial/interstadial transitions[16,17]. As can be seen from the modern and LGM distributions of *N. pachyderma* (Fig. 1), its relative abundance at ODP site 983 is sensitive to latitudinal movements of the polar front. The site is also in the general path of drifting ice originating from Iceland, Greenland and Scandinavia[4,18,19] (Extended Data Fig. 1). The IRD we identify in ODP site 983 is predominantly quartz and volcanic material (Extended Data Fig. 2), with the latter presumably sourced from Iceland[4,20] or Eastern Greenland[20] and we note that volcanic material sourced from these regions is one of the earliest arrivals within the broader episodes of ice rafting across much of the North Atlantic[4,20] (Extended Data Fig. 1). This suggests that our site is ideally positioned to detect ice-rafting events in their earliest stages.

Our results reveal the intimate association between ice rafting and high-latitude temperature variability over the last four glacial cycles with unprecedented detail (Fig. 2). The resolution of our records permits us to investigate the precise phasing between these parameters for a large number of transitions. Accordingly, we developed an algorithm for objectively assessing the temporal offsets between abrupt cooling (warming) events and the arrival (disappearance) of IRD (Methods) and we found a clear difference between episodes of cooling and warming (Fig. 3). For the majority of events, cooling (that is, an abrupt increase in %NPS) occurs before the arrival of IRD, whereas there is much closer alignment between warming and the disappearance of IRD. This result is insensitive to the choice of thresholds used to detect the transitions (Extended Data Fig. 3) and for a reasonable range of threshold values we can state that the appearance of IRD lags behind cooling for at least 75% of detected events with at least 50% of cooling events occurring more than 200 years before the arrival of icebergs. We can therefore state that if the arrival of IRD to ODP site 983 heralds the delivery of rafted ice to the broader North Atlantic[4,20] then icebergs were not the trigger for North Atlantic cold events. Occasionally, an increase in IRD may occur without a corresponding increase in %NPS (Fig. 2). This tends to happen when conditions are already cold and may reflect saturation of the %NPS proxy. Cooling events may also occur

[1]School of Earth and Ocean Sciences, Cardiff University, Cardiff CF10 3AT, UK. [2]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bussestrasse 24, D-27570 Bremerhaven, Germany. [3]Department of Geography, University College London, London WC1E 6BT, UK. [4]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. †Present address: School of Biological Sciences, University of Bristol, Bristol BS8 1TH, UK.
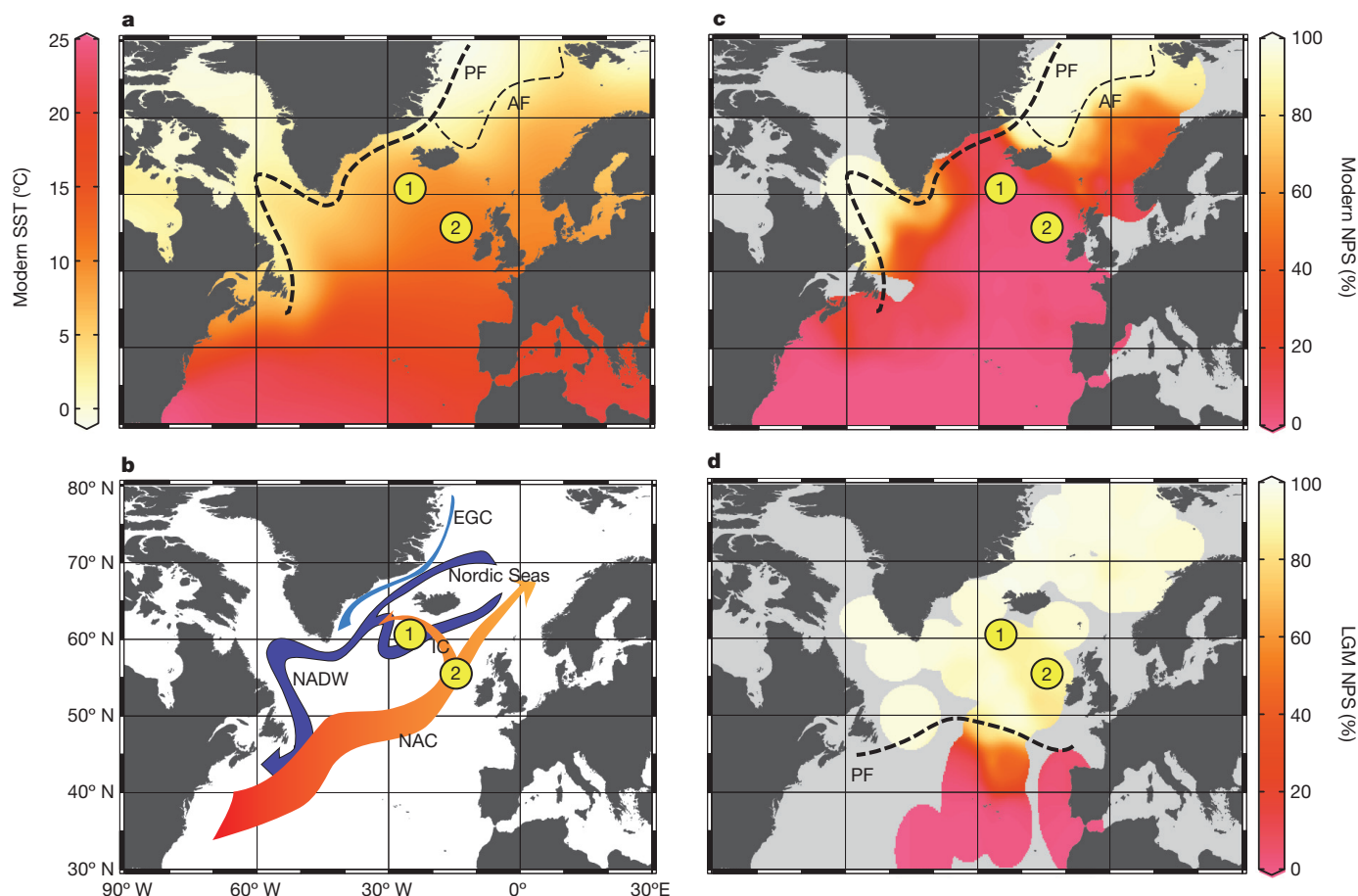
Figure 1 | Regional context of the study site. a, Modern sea surface temperature (SST)[29] shown on colour scale in degrees Celsius. PF, polar front; AF, Arctic front. **b**, Major ocean currents. NAC, North Atlantic Current; IC, Irminger Current; EGC, East Greenland Current; NADW,

North Atlantic Deep Water. **c, d**, Modern and LGM distribution of %NPS (shown on colour scale)[13]. Also shown are the locations of ODP site 983 (point 1) and ODP site 980 (point 2). The figure was generated using Ocean Data View software (http://odv.awi.de/).

without a corresponding peak in IRD. Typically, this happens earlier in a glacial cycle and may reflect the smaller size of continental ice sheets at these times. Again, this suggests that icebergs are not necessary to initiate cold events.

Notably, the asynchrony we observe between temperature and IRD at ODP site 983 does not characterize the whole of the North Atlantic. When we apply our algorithm to equivalent records from ODP site 980 (~750 km to the southeast of our site; 55.5° N, 14.7° W, 2,180 m depth; Fig. 1)[21,22] we find that both cooling and warming transitions are aligned with the appearance and disappearance of IRD, respectively (Fig. 3). The surface records from ODP sites 983 and 980 can be aligned by tuning their respective benthic $\delta^{18}O$ records (Fig. 4). Although this approach lacks precision on a millennial-timescale it is clear that cold events at ODP site 983 last longer than those at ODP site 980 and typically start earlier. Furthermore, considering the general relations between temperature and IRD depicted in Fig. 3, the most parsimonious solution is the alignment of warming between the sites. This is in line with the current consensus that abrupt warming events may be considered as essentially synchronous across the wider North Atlantic region[23] and is consistent with modelling studies using both hosing (freshwater perturbation) and non-hosing scenarios[7,16,24]. It also implies that the IRD events observed at ODP sites 983 and 980 were approximately coeval and could therefore reflect more widespread ice rafting across the North Atlantic. The observation that cooling (implied by an abrupt increase in %NPS) at ODP site 983 may occur hundreds to thousands of years earlier than at ODP site 980 is at odds with model

simulations using freshwater forcing to trigger cold events, which typically predict wholesale regional cooling within a few decades[5−7].

Instead, we suggest that the diachronous nature of cooling transitions recorded at ODP sites 983 and 980 can be explained by more gradual regional cooling and corresponding southward migration of the polar front (Fig. 4). The relative positions of ODP sites 983 and 980 means that transport of warm surface waters into the Nordic Seas could be maintained even if the polar front had migrated south of ODP site 983 (yet was still north of ODP site 980). With continued cooling the northward surface heat transport would decrease below the threshold necessary to sustain vigorous convection in the Nordic Seas (point B in Fig. 4c). At this point we suggest that the main locus of deep convection would shift to the south of Iceland as the AMOC switched from a warm to cold (stadial) mode[16]. This would coincide with a sharp increase in seasonal sea ice cover across the Nordic Seas and consequently much lower winter temperatures over Greenland as the climate entered a stadial state[25]. The abrupt transition to stadial conditions would result in rapid cooling across much of the North Atlantic[6,16], including ODP site 980 and south of the NAC (Fig. 4).

Studies suggest that the build-up of sub-surface heat in the high-latitude North Atlantic during stadials[17] may cause an increase in iceberg calving[9−11] (Methods). In combination with lower temperatures allowing wider dispersal of icebergs, this could explain the (approximately) simultaneous appearance of IRD across the wider North Atlantic at these times and suggests that the appearance of IRD at ODP site 983 is indicative of a transition to stadial conditions. This assertion is supported by the record of benthic foraminiferal $\delta^{13}C$
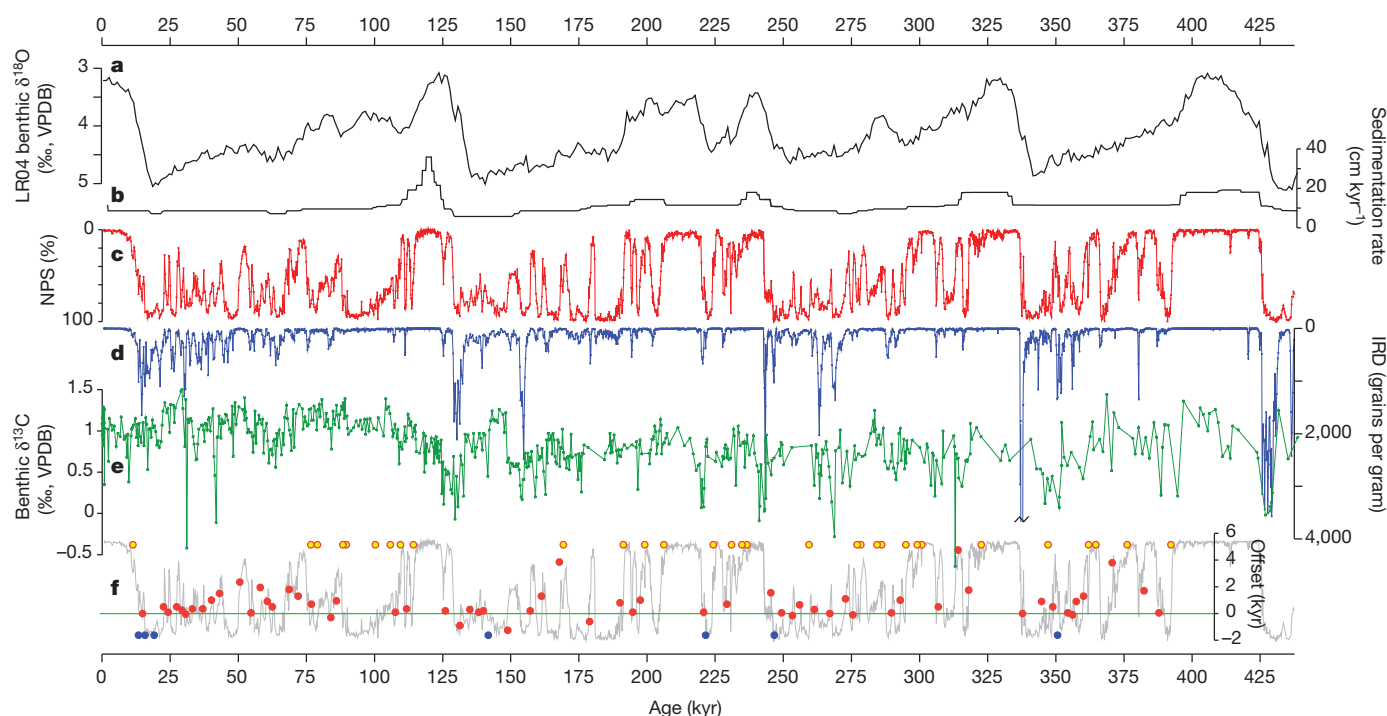
Figure 2 | Proxy records from ODP site 983. a, The LR04 benthic δ^{18}O stack[30]. **b**, Sedimentation rates in ODP site 983 according to the LR04 age model[30]. **c**, %NPS. **d**, IRD per gram. **e**, Benthic foraminiferal δ^{13}C (ref. 26). **f**, Red symbols are calculated offsets between cooling (increasing %NPS) and the arrival of IRD (a positive offset signifies that cooling occurs first). Yellow symbols are cooling events without corresponding IRD peaks and blue symbols are IRD events without registered cooling (see text for explanation). All records are plotted on the LR04 age model[30].

from ODP site 983 (ref. 26) (Fig. 2 and Extended Data Figs 4−7). Although the δ^{13}C record generally has lower temporal resolution than our records it is apparent that minima in benthic δ^{13}C tend to be shorter in duration than the cold events, as defined by %NPS, and more in line with the delivery of IRD[27]. Previous studies have interpreted low benthic δ^{13}C values at ODP site 983 to reflect increased sea ice cover over the Nordic Seas[26] or the enhanced influence of an underlying (southern-sourced) water mass with low δ^{13}C (ref. 27). Both of these conditions may be met when ocean circulation is in a stadial mode and thus the correspondence between IRD and benthic δ^{13}C implies that cooling at ODP site 983 occurs before the transition to stadial conditions.

Our observations have important chronologic implications because they suggest that a distinction should be made between stadial events *sensu stricto* (as recorded by Greenland ice cores) and North Atlantic cold events in their wider sense. Specifically, we suggest that sites to the northwest of the NAC may experience pronounced cooling (that is, a transition to Arctic/polar conditions) before the onset of Greenland stadial conditions, while those south of the NAC cool in phase with the transition to a stadial state. Indeed, a previous study from north of the NAC noted a systematic lag of 220 ± 100 years between abrupt cooling events at their site and the arrival of IRD during Marine Isotope Stage (MIS) 3 (ref. 8). On the other hand, we would expect sites throughout the North Atlantic region (both north and south of the NAC) to experience more gradual cooling before the transition to stadial conditions (as observed over Greenland[2,28]) and we note that Bond and Lotti[4] observed longer-term coolings before the arrival of IRD for several events during MIS 3. On the basis of these arguments we develop a strategy for refining the age model of ODP site 983 that can be used in future studies (see Methods and Extended Data Figs 4−8).

Our findings suggest that stadial transitions may occur as a nonlinear response to more gradual cooling, implying the existence of a threshold beyond which the transition to a stadial state becomes inevitable (Fig. 4). Indeed, the presence of such a threshold is apparent from the Greenland temperature record itself, with the duration of interstadials being a function of the rate of interstadial cooling[28] (Extended Data Fig. 9). Our results therefore support suggestions that abrupt climate transitions on millennial timescales are strongly dependent on internal feedbacks within the climate system (Methods). Increased iceberg calving and dispersal during stadials may provide a positive feedback on the AMOC, enhancing and/or prolonging stadial conditions through the addition of
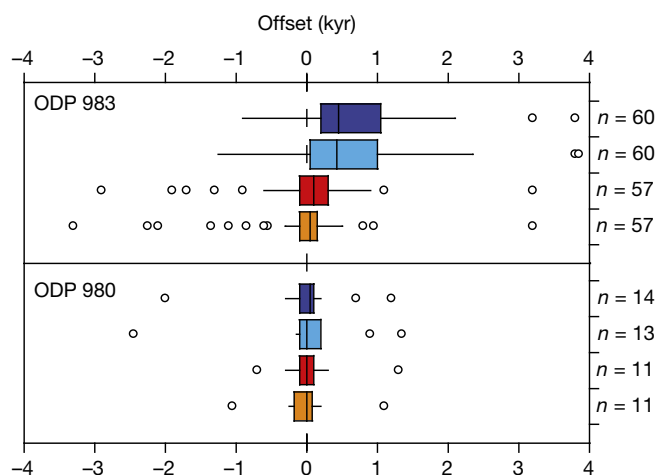
Figure 3 | Relative timing of temperature change versus ice rafting. Calculated offsets between temperature change (change in %NPS) and IRD at ODP site 983 (440−0 kyr ago) and ODP site 980 (440−360 kyr ago and 140−70 kyr ago)[21,22]. All analyses performed using the LR04 age model and identical threshold parameters (Methods). Boxes represent the interquartile range (IQR) dissected by the median value. Whiskers are 1.5 × IQR and extend to the last value included in this range. Positive values signify that temperature change is earlier. Blue boxes represent cooling versus arrival of IRD; red/orange boxes represent warming versus IRD decrease. Dark blue/red boxes represent the start of a transition; light blue/orange boxes reflect the midpoint. *n* = number of paired transitions detected.
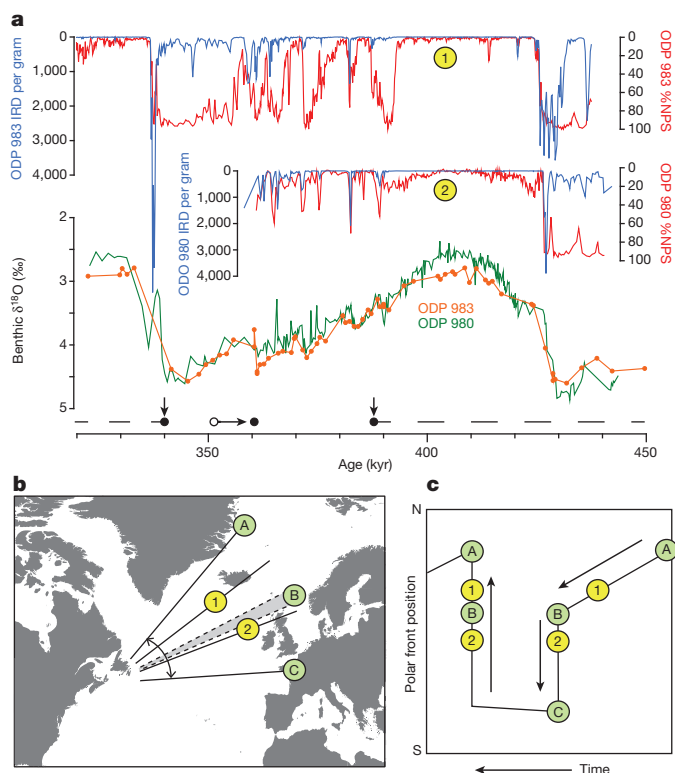
Figure 4 | Gradual cooling precedes the transition to stadial conditions.
a, Records of %NPS and IRD per gram from ODP site 983 and 980 (ref. 21) reveal earlier cooling and longer cold intervals at the northern site (Methods).
b, Cartoon showing approximate migration path of the polar front (1 and 2 are the positions of ODP sites 983 and 980). **c**, Schematic of the proposed time evolution of polar front movement. From point A, gradual cooling pushes the polar front southwards, crossing ODP site 983. On reaching threshold point B an abrupt southward migration of the polar front occurs with the transition to stadial conditions (point C). The return to warm conditions is essentially synchronous across the North Atlantic.

freshwater (with Heinrich events being the ultimate expression)[9–11,24]. However, these events should be viewed as a consequence of stadial conditions and not the driver.

Online Content Any additional Methods, Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. Science **317**, 793–796 (2007).
2. North Greenland Ice Core Project members. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. Nature **431**, 147–151 (2004).
3. Broecker, W. S., Peteet, D. M. & Rind, D. Does the ocean-atmosphere system have more than one stable mode of operation? Nature **315**, 21–26 (1985).
4. Bond, G. C. & Lotti, R. Iceberg discharges into the North Atlantic on millennial time scales during the last glaciation. Science **267**, 1005–1010 (1995).
5. Rahmstorf, S. Rapid climate transitions in a coupled ocean-atmosphere model. Nature **372**, 82–85 (1994).
6. Vellinga, M. & Wood, R. A. Global climatic impacts of a collapse of the Atlantic thermohaline circulation. Clim. Change **54**, 251–267 (2002).
7. Kageyama, M. et al. Climatic impacts of fresh water hosing under Last Glacial Maximum conditions: a multi-model study. Clim. Past **9**, 935–953 (2013).
8. van Kreveld, S. et al. Potential links between surging ice sheets, circulation changes, and the Dansgaard-Oeschger cycles in the Irminger Sea, 60–18 kyr. Paleoceanography **15**, 425–442 (2000).
9. Shaffer, G., Olsen, S. M. & Bjerrum, C. J. Ocean subsurface warming as a mechanism for coupling Dansgaard-Oeschger climate cycles and ice-rafting events. Geophys. Res. Lett. **31**, L24202 (2004).
10. Clark, P. U., Hostetler, S. W., Pisias, N. G., Schmittner, A. & Meissner, K. J. Mechanisms for an 7-kyr climate and sea-level oscillation during Marine Isotope Stage 3. Ocean Circulation: Mechanisms and Impacts—Past and Future Changes of Meridional Overturning (eds Schmittner, A. Chiang, J. C. H. & Hemming, S. R.) 209–246 (AGU, 2007).
11. Alvarez-Solas, J., Robinson, A., Montoya, M. & Ritz, C. Iceberg discharges of the last glacial period driven by oceanic circulation changes. Proc. Natl Acad. Sci. USA **110**, 16350–16354 (2013).
12. Dickson, R. R. & Brown, J. The production of North Atlantic DeepWater—sources, rates, and pathways. J. Geophys. Res. Oceans **99**, 12319–12341 (1994).
13. MARGO Project members. Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum. Nature Geosci. **2**, 127–132 (2009).
14. Sarnthein, M. et al. Changes in East Atlantic deep-water circulation over the last 30,000 years: eight time slice reconstructions. Paleoceanography **9**, 209–267 (1994).
15. Weber, S. L. et al. The modern and glacial overturning circulation in the Atlantic Ocean in PMIP coupled model simulations. Clim. Past **3**, 51–64 (2007).
16. Ganopolski, A. & Rahmstorf, S. Rapid changes of glacial climate simulated in a coupled climate model. Nature **409**, 153–158 (2001).
17. Ezat, M. M., Rasmussen, T. L. & Groeneveld, J. Persistent intermediate water warming during cold stadials in the southeastern Nordic seas during the past 65 k.y. Geology **42**, 663–666 (2014).
18. Death, R., Siegert, M. J., Bigg, G. R. & Wadley, M. R. Modelling iceberg trajectories, sedimentation rates and meltwater input to the ocean from the Eurasian Ice Sheet at the Last Glacial Maximum. Palaeogeogr. Palaeoclimatol. Palaeoecol. **236**, 135–150 (2006).
19. Watkins, S., Maher, B. & Bigg, G. Ocean circulation at the Last Glacial Maximum: a combined modeling and magnetic proxy-based study. Paleoceanography **22**, PA2204 (2007).
20. Jullien, E. et al. Contrasting conditions preceding MIS3 and MIS2 Heinrich events. Global Planet. Change **54**, 225–238 (2006).
21. Oppo, D. W., McManus, J. F. & Cullen, J. L. Abrupt climate events 500,000 to 340,000 years ago: evidence from subpolar north Atlantic sediments. Science **279**, 1335–1338 (1998).
22. Oppo, D. W., McManus, J. F. & Cullen, J. L. Evolution and demise of the Last Interglacial warmth in the subpolar North Atlantic. Quat. Sci. Rev. **25**, 3268–3277 (2006).
23. Austin, W. E. & Hibbert, F. D. Tracing time in the ocean: a brief review of chronological constraints (60–8 kyr) on North Atlantic marine event-based stratigraphies. Quat. Sci. Rev. **36**, 28–37 (2012).
24. Knorr, G. & Lohmann, G. Rapid transitions in the Atlantic thermohaline circulation triggered by global warming and meltwater during the last deglaciation. Geochem. Geophys. Geosyst. **8**, Q12006 (2007).
25. Li, C., Battisti, D. S. & Bitz, C. M. Can North Atlantic sea ice anomalies account for Dansgaard-Oeschger climate signals? J. Clim. **23**, 5457–5475 (2010).
26. Raymo, M. E. et al. Stability of North Atlantic water masses in face of pronounced climate variability during the Pleistocene. Paleoceanography **19**, PA2008 (2004).
27. Kleiven, H. F., Hall, I. R., McCave, I. N., Knorr, G. & Jansen, E. Coupled deep-water flow and climate variability in the middle Pleistocene North Atlantic. Geology **39**, 343–346 (2011).
28. Schulz, M. The tempo of climate change during Dansgaard-Oeschger interstadials and its potential to affect the manifestation of the 1470-year climate cycle. Geophys. Res. Lett. **29**, http://dx.doi.org/10.1029/2001GL013277(2002).
29. Locarnini, R. A. et al. in NOAA Atlas NESDIS 68 (ed. Levitus, S.) 184 (US Government Printing Office, 2010).
30. Lisiecki, L. E. & Raymo, M. E. A. Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. Paleoceanography **20**, http://dx.doi.org/10.1029/2004PA001071 (2005).

Author Contributions S.B. designed research and analysed datasets. J.C. processed samples and performed faunal counts with assistance from those mentioned in the Acknowledgements. X.G. performed ice core data analysis. S.B., L.J., X.G., G.K. and D.T. contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.B. (barkers3@cf.ac.uk).

METHODS

Sample preparation and faunal counts. Sediment samples were spun overnight and washed with deionized water through a 63-μm sieve before being dried at 40 °C. IRD and faunal counts were made on the >150 μm fraction after splitting to yield approximately 300 entities. Only left-coiling specimens of *N. pachyderma* were counted and those with morphological resemblance to *Neogloboquadrina incompta* were not counted. The uncertainty due to aberrant coiling in *N. pachyderma* is therefore <3% (ref. 31). The percentage of *N. pachyderma* in the North Atlantic can be used as a sensitive tracer for the locations of oceanic fronts in this region (Fig. 1). According to Pflaumann *et al.*[32] the Arctic front is documented by the transition from ~90%−94% NPS, while values of ~98% NPS track the polar front. IRD was considered to be the total number of lithogenic/terrigenous grains counted. The majority of grains fall into two categories: quartz and volcanics, with volcanics comprising ~36% of the total IRD on average (Extended Data Fig. 2).

Temporal offsets between temperature change and IRD input. *Code availability.* Temporal offsets were determined using an algorithm developed in Matlab (the script is available as a Matlab file in the Supplementary Information).

All datasets were input in the time domain (equivalent results were obtained using the depth domain) using the LR04 timescale[30] (a revised age model was also used for comparison, Extended Data Fig. 8) and evenly resampled at 0.1 kyr intervals (similar to the physical sampling rate during interglacial periods, when sedimentation rates are greatest). Records were then smoothed using a rectangular filter (running mean) of 0.5 kyr (similar to the lowest sampling rate during full glacial periods) implemented by *filtfilt* in Matlab (that is, run forward and reverse) and differentiated with respect to time (via the difference quotient). Abrupt transitions in %NPS or IRD per gram then identified by their respective derivatives exceeding a threshold. When looking for cooling events, the algorithm is primed by the completion of a warming event according to % NPS (completion of an IRD event serves as an alternative primer). It then searches for the next time %NPS and/or IRD per gram increases at a rate greater than a given threshold (specific to each parameter). The algorithm is reset when warming next occurs. Warming offsets (decrease in %NPS and IRD per gram) are quantified in an analogous way with a threshold value equal to −1 times that used for the cooling offsets. The algorithm identifies the onset of a transition as the time when the threshold is first exceeded and the mid-point of the transition as the mid-point of all consecutive points exceeding the threshold. We calculate offsets for both the start and mid-point of transitions since the mid-point is less sensitive to the specific threshold values employed. However, we find a similar result using either approach (Fig. 3). The algorithm rejects offsets outside of a given range, in this case ±6 kyr.

The detection of individual events depends on a trade-off between the length of smoothing window applied (which is common to all records) and the derivative threshold values employed (which are specific to each record). To determine the optimal set of threshold values we performed a sensitivity analysis (Extended Data Fig. 3). If a threshold is too sensitive or too insensitive it is less likely that a true pair of transitions will be identified, leading to an erroneous calculated offset. This effect is apparent from offsets calculated for warming events. It can be seen that the interquartile range (IQR) for warming offsets is larger when the thresholds are at the lower or upper limits of our sensitivity analysis. An equivalent result is obtained when IRD is used as a primer. We suggest that the most appropriate threshold pairs should result in smaller values of the IQR for warming transitions (implying greater consistency between individual events). From the results shown in Extended Data Fig. 3 we use this principle to delineate a region of optimal threshold values. We employ the lower left set of values within this range in Fig. 3 because it results in the highest number of paired transitions. It is also the most conservative in terms of the calculated offsets (other pairings result in more positive offsets).

Alignment of records from ODP site 983 and ODP site 980. See Fig. 4. The cores were aligned by modifying the LR04 age model[30] to improve alignment of their benthic δ18O records. In Fig. 4, records from ODP site 980 are on LR04 throughout. Those from ODP site 983 are on LR04 for the dashed intervals. Between these anchor points the age model for ODP site 983 has been shifted as indicated by the horizontal black arrow.

Stadial transitions as a nonlinear response to gradual cooling. Since Bond and Lotti's[4] observation that most Greenland stadial events (not just those related to Heinrich events) were associated with increased iceberg calving and dispersal across the North Atlantic region, a great number of ocean and climate modelling studies have employed freshwater forcing scenarios to simulate abrupt change across the North Atlantic region and beyond[5−7,33,34]. Clearly the influence of freshwater on the efficacy of NADW formation has important consequences for

the AMOC, and mechanisms that invoke (quasi-) periodic fluctuations in iceberg calving and freshwater input provide an appealing solution to the question of why Greenland stadials (and North Atlantic Heinrich events) occur with such regularity[35−37]. However, the observations that cooling—both abrupt (this study and ref. 8) and gradual (ref. 4)—across the North Atlantic actually precedes the transition to stadial conditions (and therefore the release of icebergs) suggests that an alternative to freshwater forcing from icebergs should be considered as the trigger for inducing stadial transitions. Accordingly, several previous studies have invoked abrupt transitions in ocean circulation without calling on iceberg discharge as a primary forcing agent[10,38−44].

On the basis of our observations we invoke gradual cooling across the North Atlantic region as the ultimate trigger for the transition to stadial conditions. Thus we consider stadial transitions as an abrupt, nonlinear response to more gradual forcing. The precise cause(s) of cooling may be manifold and may vary depending on the background state. For example, longer-term cooling may be the result of changes in insolation, greenhouse gas forcing or ice sheet configuration[44]. On shorter timescales cooling could be induced by a gradual weakening of the AMOC either in response to a gradual freshening of the surface North Atlantic[38] or following a transient AMOC overshoot at the onset of interstadial conditions[24,33]. Alternatively, the build-up of circum-North Atlantic ice shelves could lead eventually to runaway cooling and the development of stadial conditions[43]. Once in a stadial state, the build-up of subsurface heat may lead to the destruction of ice shelves and ultimately the partial collapse of land-based ice sheets (with Heinrich events being the ultimate expression of such a mechanism)[4,9−11,43,45,46]. The freshwater provided as a consequence of such a collapse may be expected to enhance and /or prolong stadial conditions[10,24] and thus should be considered as a positive feedback rather than the initial trigger for stadial transitions.

Revised age model development. The site of ODP site 983 is positioned on the rapidly accumulating Gardar Drift and sediment accumulation is sensitive to changes in the dense overflows crossing the Iceland−Scotland Ridge[26,27], which themselves are thought to vary in concert with high-latitude climate[17,27]. At orbital timescales this can be seen through elevated sedimentation rates during interglacials (as implied by the LR04 age model[30]; Fig. 2), when the overflows are thought to be more vigorous[26,27], but this also implies that sedimentation rates are elevated during millennial-scale warm events that are not accounted for by the LR04 age model. Given the potentially large and frequent changes in sedimentation rate at ODP site 983 we require a more detailed tuning strategy for refining the age assignment of abrupt events within our records. A typical approach in the development of such an age model is to align abrupt changes in our records with those in a reference stratigraphy[23] such as GL_T_syn (a synthetic prediction of Greenland temperature)[47,48].

In line with previous studies[23,47] we assume that abrupt warming events in our record (which also align with the disappearance of IRD) are synchronous with warming across the wider North Atlantic region and align these with warming transitions in GL_T_syn[48] (Extended Data Figs 4−7). As mentioned, sedimentation at ODP site 983 is sensitive to the overflows crossing the Iceland−Scotland ridge[26,27], with elevated accumulation rates during warm intervals reflecting enhanced advection of fine (<63 μm) material to the site of ODP site 983 driven by faster currents crossing the Iceland−Faeroe ridge[26]. The coarse (>63 μm) fraction of ODP site 983 therefore reflects both the delivery of IRD (which increases during stadials) and the input of fine fraction (which decreases during stadials). We therefore align increases in the coarse fraction with cooling transitions in GL_T_syn. We tune our records to GL_T_syn on the EDC3 age model[49] and note that the implied changes in sedimentation rate are in line with expectations (higher during warmer intervals). We also convert the EDC3 ages to an alternative ice core age model (AICC2012[50,51]) and an absolute age model (GICC05/NALPS/China) based on previous studies[48,52−55] (see the source data associated with Extended Data Fig. 4).

Given the potential influence of the sedimentation rate changes implied by our new age model on the calculated offsets, we ran the algorithm on the data sets using the new age model (Extended Data Fig. 8). We note that the distribution of offsets is very similar for the two age models (LR04 versus our EDC3) although the median cooling offset (mid-transition) for the revised age model is slightly smaller at 350 years compared with 425 years when using LR04. This can be explained as a result of higher implied sedimentation rates (fewer years per sampled interval) during interstadial periods that extend beyond cooling (according to %NPS) until the arrival of IRD.

31. Darling, K. F., Kucera, M., Kroon, D. & Wade, C. M. A resolution for the coiling direction paradox in *Neogloboquadrina pachyderma*. *Paleoceanography* **21**, PA2011 (2006).

32. Pflaumann, U., Duprat, J., Pujol, C. & Labeyrie, L. D. SIMMAX: a modern analog technique to deduce Atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments. *Paleoceanography* **11,** 15−35 (1996).
33. Liu, Z. *et al.* Transient simulation of last deglaciation with a new mechanism for Bolling-Allerod warming. *Science* **325,** 310−314 (2009).
34. Menviel, L., Timmermann, A., Friedrich, T. & England, M. Hindcasting the continuum of Dansgaard−Oeschger variability: mechanisms, patterns and timing. *Clim. Past* **10,** 63−77 (2014).
35. MacAyeal, D. R. Binge/purge oscillations of the Laurentide ice sheet as a cause of the North Atlantic Heinrich events. *Paleoceanography* **8,** 775−784 (1993).
36. Alley, R. B., Anandakrishnan, S. & Jung, P. Stochastic resonance in the North Atlantic. *Paleoceanography* **16,** 190−198 (2001).
37. Braun, H. *et al.* Possible solar origin of the 1,470-year glacial climate cycle demonstrated in a coupled model. *Nature* **438,** 208−211 (2005).
38. Broecker, W. S., Bond, G. & Klas, M. A salt oscillator in the glacial Atlantic? 1. The concept. *Paleoceanography* **5,** 469−477 (1990).
39. Winton, M. & Sarachik, E. S. Thermohaline oscillations induced by strong steady salinity forcing of ocean general-circulation models. *J. Phys. Oceanogr.* **23,** 1389−1410 (1993).
40. Winton, M. The effect of cold climate upon North Atlantic Deep Water formation in a simple ocean-atmosphere model. *J. Clim.* **10,** 37−51 (1997).
41. Schulz, M., Paul, A. & Timmermann, A. Relaxation oscillators in concert: a framework for climate change at millennial timescales during the late Pleistocene. *Geophys. Res. Lett.* **29,** http://dx.doi.org/10.1029/2002GL016144 (2002).
42. Dokken, T. M., Nisancioglu, K. H., Li, C., Battisti, D. S. & Kissel, C. Dansgaard-Oeschger cycles: interactions between ocean and sea ice intrinsic to the Nordic seas. *Paleoceanography* **28,** 491−502 (2013).
43. Petersen, S., Schrag, D. & Clark, P. U. A new mechanism for Dansgaard-Oeschger cycles. *Paleoceanography* **28,** 24−30 (2013).
44. Zhang, X., Lohmann, G., Knorr, G. & Purcell, C. Abrupt glacial climate shifts controlled by ice sheet changes. *Nature* **512,** 290−294 (2014).
45. Alvarez-Solas, J. *et al.* Links between ocean temperature and iceberg discharge during Heinrich events. *Nature Geosci.* **3,** 122−126 (2010).
46. Marcott, S. A. *et al.* Ice-shelf collapse from subsurface warming as a trigger for Heinrich events. *Proc. Natl Acad. Sci. USA* **108,** 13415−13419 (2011).
47. Hodell, D. *et al.* Response of Iberian Margin sediments to orbital and suborbital forcing over the past 420 ka. *Paleoceanography* **28,** 185−199 (2013).
48. Barker, S. *et al.* 800,000 years of abrupt climate variability. *Science* **334,** 347−351 (2011).
49. Parrenin, F. *et al.* The EDC3 chronology for the EPICA dome C ice core. *Clim. Past* **3,** 485−497 (2007).
50. Bazin, L. *et al.* An optimized multi-proxy, multi-site Antarctic ice and gas orbital chronology (AICC2012): 120-800 ka. *Clim. Past* **9,** 1715−1731 (2013).
51. Veres, D. *et al.* The Antarctic ice core chronology (AICC2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. *Clim. Past* **9,.** 1733−1748 (2013).
52. Barker, S. & Diz, P. Timing of the descent into the last ice age determined by the bipolar seesaw. *Paleoceanography,* **29,** http://dx.doi.org/10.1002/2014PA002623 (2014).
53. Andersen, K. K. *et al.* A 60 000 year Greenland stratigraphic ice core chronology. *Clim. Past Discuss.* **3,** 1235− **1260,**(2007).
54. Boch, R. *et al.* NALPS: a precisely dated European climate record 120-60 ka. *Clim. Past* **7,** 1247−1259 (2011).
55. Cheng, H. *et al.* Ice age terminations. *Science* **326,** 248−252 (2009).
56. Ruddiman, W. F. Late Quaternary deposition of ice-rafted sand in the subpolar North Atlantic (lat 40 to 65 N). *Geol. Soc. Am. Bull.* **88,** 1813−1827 (1977).
57. Elliot, M. *et al.* Millennial-scale iceberg discharges in the Irminger Basin during the last glacial period: relationship with the Heinrich events and environmental settings. *Paleoceanography* **13,** 433−446 (1998).
58. Revel, M., Cremer, M., Grousset, F. E. & Labeyrie, L. Grain-size and Sr-Nd isotopes as tracer of paleo-bottom current strength, Northeast Atlantic Ocean. *Mar. Geol.* **131,** 233−249 (1996).
59. Moros, M. *et al.* Quartz content and the quartz-to-plagioclase ratio determined by X-ray diffraction: a proxy for ice rafting in the northern North Atlantic? Earth Planet. *Sci. Lett.* **218,** 389−401 (2004).
60. Prins, M. A. *et al.* Ocean circulation and iceberg discharge in the glacial North Atlantic: inferences from unmixing of sediment size distributions. *Geology* **30,** 555−558 (2002).
61. Jonkers, L. *et al.* A reconstruction of sea surface warming in the northern North Atlantic during MIS 3 ice-rafting events. *Quat. Sci. Rev.* **29,** 1791−1800 (2010).
62. Scourse, J. D., Hall, I. R., McCave, I. N., Young, J. R. & Sugdon, C. The origin of Heinrich layers: evidence from H2 for European precursor events. *Earth Planet. Sci. Lett.* **182,** 187−195 (2000).
63. Grousset, F. E., Pujol, C., Labeyrie, L., Auffret, G. & Boelaert, A. Were the North Atlantic Heinrich events triggered by the behavior of the European ice sheets? *Geology* **28,** 123−126 (2000).
64. Grousset, F. E. *et al.* Zooming in on Heinrich layers. *Paleoceanography* **16,** 240−259 (2001).

LETTER

A Mercury–like component of early Earth yields uranium in the core and high mantle ^{142}Nd

Anke Wohlers[1] & Bernard J. Wood[1]

Recent ^{142}Nd isotope data indicate that the silicate Earth (its crust plus the mantle) has a samarium to neodymium elemental ratio (Sm/Nd) that is greater than that of the supposed chondritic building blocks of the planet. This elevated Sm/Nd has been ascribed either to a 'hidden' reservoir in the Earth[1,2] or to loss of an early-formed terrestrial crust by impact ablation[3]. Since removal of crust by ablation would also remove the heat-producing elements—potassium, uranium and thorium—such removal would make it extremely difficult to balance terrestrial heat production with the observed heat flow[3]. In the 'hidden' reservoir alternative, a complementary low-Sm/Nd layer is usually considered to reside unobserved in the silicate lower mantle. We have previously shown, however, that the core is a likely reservoir for some lithophile elements such as niobium[4]. We therefore address the question of whether core formation could have fractionated Nd from Sm and also acted as a sink for heat-producing elements. We show here that addition of a reduced Mercury-like body (or, alternatively, an enstatite-chondrite-like body) rich in sulfur to the early Earth would generate a superchondritic Sm/Nd in the mantle and an ^{142}Nd/^{144}Nd anomaly of approximately +14 parts per million relative to chondrite. In addition, the sulfur-rich core would partition uranium strongly and thorium slightly, supplying a substantial part of the 'missing' heat source for the geodynamo.

Terrestrial rocks were recently found to have higher ratios of radiogenic ^{142}Nd to nonradiogenic ^{144}Nd than do the chondritic meteorites generally supposed to be representative of the material from which Earth accreted[1,2]. ^{142}Nd was produced during the early history of the Solar System from decay of the extinct radionuclide ^{146}Sm (half-life, $t_{1/2} = 68$ million years[5] and the presence of a positive ^{142}Nd anomaly of ~20 parts per million (p.p.m.) calculated as $10^6\left[\left(\frac{^{142}Nd}{^{144}Nd}\right)_{Earth} - \left(\frac{^{142}Nd}{^{144}Nd}\right)_{chondrite}\right] / \left(\frac{^{142}Nd}{^{144}Nd}\right)_{Earth}$) or of ~9 p.p.m.[6] in the silicate Earth would require an Sm/Nd ratio higher than chondritic[1,2]. This high Sm/Nd ratio was established early in Earth's history while ^{146}Sm was still 'alive' (that is, undergoing radioactive decay).

A plausible mechanism for generating high Sm/Nd in Earth's mantle is partial melting and melt extraction to form a crust. Because Nd is less compatible in mantle silicates than Sm[7], partial melts have relatively low Sm/Nd and the solid residue has high Sm/Nd. A low-Sm/Nd crust could be completely removed from the mantle system by subduction to an inaccessible region of the deep mantle[1] or removed from Earth by impact ablation[3]. The problem with the former hypothesis is the lack of evidence for a hidden silicate reservoir, while the latter hypothesis suffers from the requirement that much of Earth's heat production, in the form of radioactive uranium (U), thorium (Th) and potassium (K) would be removed together with the low-Sm/Nd crust. Assuming chondritic abundances of U and Th and a K/U ratio of ~12,000 for the silicate Earth[8], the heat production in the Earth is only about 0.6 times the current heat loss[9]. Reducing the heat sources further by ablation loss would make it even more difficult to reconcile heat production with heat loss.

An additional question in the context of heat production is that of the energy source for the Earth's magnetic field[10]. Arising from convection in the core, Earth has had a magnetic field for at least 3.5 billion years. The crystallization of the inner core is an important source of energy for the geodynamo[11] but most attempts to construct histories of core cooling indicate that the inner core cannot be much older than $1-1.5$ billion years[10,11] unless a source of radioactive heating is present. Numerous studies have focused on ^{40}K as a potential core heat source, because K, in common with all moderately volatile elements[8], is depleted in the silicate Earth relative to the chondritic abundance.

Furthermore, high-pressure experiments[12,13] indicate that K enters sulfide under oxidizing conditions and sulfur (S) is believed to be a major component of the core's complement of approximately 10% of elements of low atomic number[14]. It appears, however that the maximum possible K content of the core is insufficient to generate more than a small fraction of the $2-5$ TW required to generate reasonable core thermal histories[11,13]. The alternative explanation—that U and/or Th provide the energy for core convection—has some support from early experiments on sulphide−silicate partitioning[15] but more recent results indicate very little partitioning of U into S-bearing metals even under extreme conditions[16].

We approached the problem of U, Th, Nd and Sm in Earth's mantle and core from the standpoint of recent work on partitioning between sulfide melts and silicate melts[17]. Kiseeva and Wood[17] found that the sulfide−silicate partition coefficient for any element i, defined as $D_i = [i]_{sulf}/[i]_{sil}$, is dependent on the FeO content of the silicate melt, such that for FeS-rich sulfides:

$$\log D_i = A - n\log[FeO] \tag{1}$$

where A is a constant and n is a constant dependent on the valency of element i. Therefore, under strongly reducing conditions, where the FeO content of the silicate melt is very low (<1% for example) one would expect D_i values to be much higher than under the conditions of MORB crystallization where the FeO content of the melt is about 8%−10%. This hypothesis is consistent with the data of Murrell and Burnett[15], who observed strong partitioning of U into sulfide liquid at low oxygen fugacity f_{O_2}. Given terrestrial accretion models calling for prolonged periods of growth under reduced conditions[18,19], the demonstration that Mercury is a highly reduced S-rich planet with a liquid core[20,21] and the association of the rare earth elements (REEs), U and Th with sulfides in enstatite chondrite meteorites[22], we investigated partitioning of U, Th, Sm, Nd and several other lithophile elements into liquid iron sulfide under reducing conditions.

Experiments were performed at 1.5 GPa and temperatures between 1,400 °C and 1,650 °C using starting materials that were approximately 50:50 mixtures of silicate and FeS doped with a range of lithophile trace elements including U, Th, La, Nd, Sm, Eu, Yb, Ce, and Zr (see Methods). The silicate was a basalt-like composition in the system $CaO−MgO−Al_2O_3−SiO_2$ with variable FeO. Analysis

[1]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK.

was by electron microprobe and laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) (Methods). Table 1 presents a summary of sulfide−silicate partitioning results (see Extended Data Figs 1−3 and Extended Data Tables 1−4 for complete analyses).

Figure 1a shows data from a series of experiments performed at 1,400 °C. As can be seen, the partition coefficients of U, Nd and Sm are strong functions of the FeO content of the silicate melt, increasing dramatically, as predicted, as the FeO content decreases below 1 wt%. The negative slope of $\log D_i$ as $f(\log[\text{FeO}_{sil}])$ reverses at high FeO_{sil}, however, because the sulfide dissolves progressively more oxygen as the FeO content of silicate increases and these three lithophile elements (Fig. 1a) follow oxygen into the sulfide. We found similar behaviour in two more series of experiments at higher temperature (Table 1 and Extended Data Fig. 1). Other lithophile elements, notably Ti, Nb and Ta (B.J.W., unpublished data) behave similarly. Importantly, we find $D_U > D_{Nd} > D_{Sm}$ for partitioning into sulfide in all experiments. At very low FeO contents all D_i become >1 (Fig. 1a). Furthermore (Fig. 1b) D_{Nd} is always appreciably greater than D_{Sm}, with D_{Nd}/D_{Sm} approaching 1.5 in some cases.

The implications of Fig. 1 are that segregation of sulfide (or S-rich metal) from reduced FeO-poor silicate will lead to enrichment of the metallic phase in U and in Nd relative to Sm when compared to the silicate. Addition of such material to the core and mantle respectively of a growing planet would provide a core heat source and a mantle with superchondritic Sm/Nd, Yb/Sm and Yb/La. Although potentially detectable in terms of a mantle [142]Nd anomaly, the fractionation of heavy from light REEs (Table 1) in the primitive mantle of the body (the bulk silicate Earth in this case) would have little effect on its overall REE pattern (Extended Data Fig. 3). Similarly, there would be no observable Eu anomaly despite the fact that Eu is probably in the 2^+ oxidation state (unlike the other 3^+ lanthanides) under these conditions (Extended Data Fig. 3). If such a body represented Earth early in its history then the mantle would have a positive [142]Nd anomaly relative to chondrite (as observed) and much of the energy deficit identified for core convection[10] would be supplied by U (and Th). We find that D_{Th}/D_U is about 0.1, indicating that U would be accompanied by Th in the S-rich core. Addition of more-oxidized material later in accretion would lead to the higher current FeO content of the mantle (8.1%)[8], but could not erase the superchondritic Sm/Nd ratio of the mantle and U content of the core unless there were complete core−mantle re-equilibration.

Figure 2 illustrates the impact of adding a highly reduced body rich in sulfide to the growing Earth. The Th/U ratio of the silicate Earth would be higher than chondritic (3.8−3.9[8,23]), which provides an important constraint on how much U can be present in Earth's core. Based on the Pb-isotopic compositions of Archean galenas[24] and of 3.5-billion-year-old komatiites[25] the Th/U ratio of the Archean mantle has been estimated to be ≥4.3. Tatsumoto[26] argued, on the basis of the Pb isotopic compositions of basalts, for an early differentiation of the mantle, which resulted in a Th/U of 4.2−4.5 in the mantle source regions. Since that time the Th/U ratio of the mantle has decreased, probably owing to preferential recycling of the more soluble U[27].

Figure 2 shows four models of U content of the core and the [142]Nd anomaly of the mantle (relative to the bulk Earth), based on our partitioning data. We choose a reduced body of 0.15 mass fraction sulfide, corresponding to the S content of primitive CI chondrites[8] and use values of D_{Sm} ($\text{Sm}_{sulf}/\text{Sm}_{sil}$) that are close to the observed maximum of 0.8−2.2, noting that D_{Sm} values increase with decreasing temperature and that segregation of sulfide from a crystal-melt mush instead of melt alone would increase them further because of the incompatibility of Sm, Nd and U in crystals. As can be seen (Fig. 2a), adding 20% of such a body to Earth would lead to 4−5 parts per billion (p.p.b.) of U in the core, a Th/U of the silicate Earth of 4.17 and a [142]Nd anomaly in the mantle relative to the bulk Earth
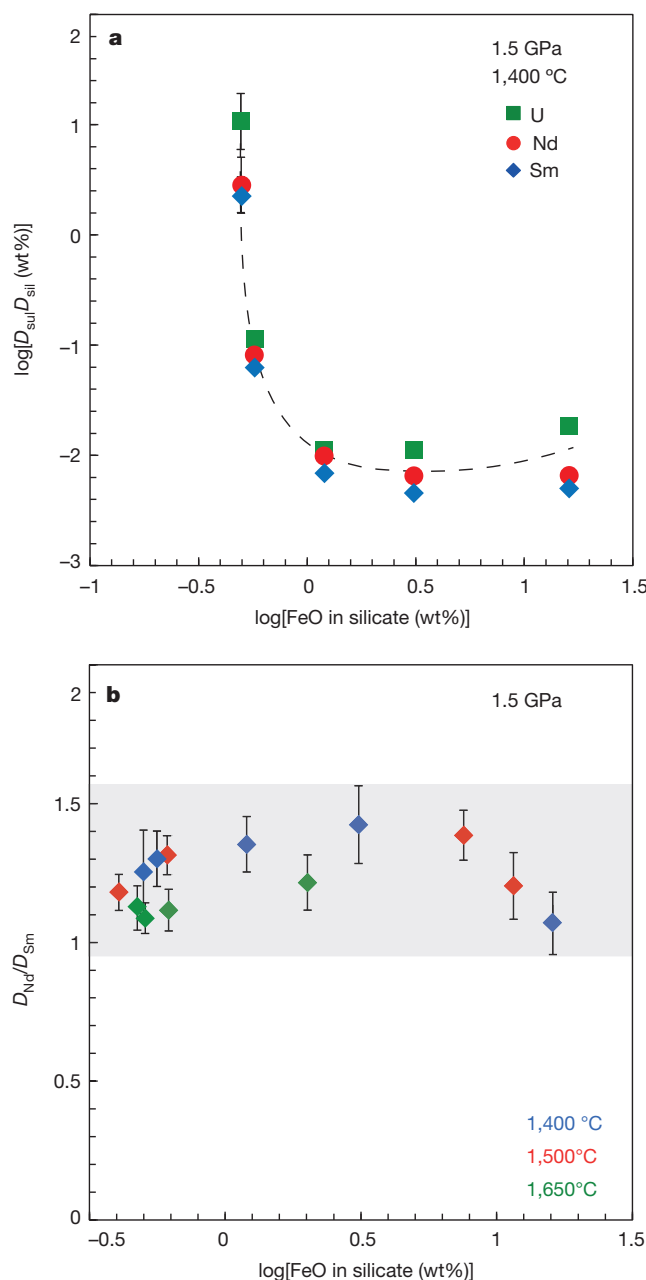
Figure 1 | Sulphide−silicate partitioning data. a, Partition coefficients $D_i = [i]_{sulf}/[i]_{sil}$ for U, Nd and Sm at 1.5 GPa and 1,400 °C plotted versus the log of the FeO content of the silicate melt in weight per cent. **b**, The ratio of D_{Nd}/D_{Sm} plotted versus log[FeO]. Error bars in both cases are ±2 s.e. and, if absent, are smaller than symbol size.

of ~7 p.p.m. Increasing the reduced body mass to 45% (Fig. 2b) leads to about 8 p.p.b. U in the core, a Th/U of the silicate Earth of 4.5 and a mantle [142]Nd anomaly of 13.9 p.p.m. relative to the bulk Earth.

We performed a sensitivity analysis (Extended Data Fig. 2) and find that, if the Th/U of the silicate Earth is ≤4.5, the maximum U content of the core is 8 p.p.b. with a Th content of ~8 p.p.b. These figures increase to ~10 p.p.b. if the Th/U of the silicate Earth is ≤4.7. The [142]Nd anomaly is 13.9 p.p.m. in the former case and ~17 p.p.m. in the latter. The estimated U and Th contents of the core would lead to 2−2.4 TW, sufficient to power the geodynamo[11] even without the potential 0.4−0.8 TW from [40]K decay[13]. We can reduce the size of the reduced body by increasing its S content (Fig. 2c, d) and increase D_U/D_{Sm} but the overall effects on the core and mantle [142]Nd remain

Table 1 | Summary of sulfide−silicate partition coefficients

Run	Pressure, P (GPa)	Temperature, T (°C)	log[FeO_{sil}(wt%)]	D_{Sm}	D_{Nd}/D_{Sm}	D_U/D_{Sm}	D_{Th}/D_U	D_{Eu}/D_{Sm}	D_{La}/D_{Sm}	D_{Yb}/D_{Sm}
421	1.5	1,400	0.50	0.005	1.42	2.47	0.036	5.85	1.38	0.16
σ				0.001	0.29	0.37	0.008	0.93	0.23	0.07
428	1.5	1,400	0.08	0.013	1.35	1.56	0.038	5.50	1.35	0.16
σ				0.001	0.19	0.17	0.005	1.06	0.20	0.02
427	1.5	1,400	−0.25	0.062	1.30	1.81	0.028	2.36	1.23	0.13
σ				0.006	0.19	0.24	0.004	0.40	0.22	0.02
426	1.5	1,400	−0.30	2.247	1.25	6.81	0.048	0.14	1.03	0.16
σ				0.333	0.30	1.43	0.010	0.05	0.15	0.09
429	1.5	1,400	1.21	0.005	1.04	3.68	0.200	2.04	1.10	0.67
σ				0.0001	0.12	0.89	0.041	0.42	0.47	0.19
461	1.5	1,650	0.30	0.023	1.22	1.92	0.058	4.09	1.18	0.21
σ				0.003	0.20	0.32	0.009	0.58	0.18	0.03
462	1.5	1,650	−0.21	0.154	1.12	3.58	0.046	1.13	0.92	0.27
σ				0.011	0.13	0.43	0.007	0.14	0.10	0.03
477	1.5	1,650	−0.29	0.629	1.10	9.26	0.044	0.37	0.83	0.39
σ				0.038	0.11	0.73	0.043	0.38	0.80	0.39
464	1.5	1,650	−0.32	0.751	1.13	9.41	0.035	0.21	0.55	0.28
σ				0.073	0.16	1.18	0.020	0.16	0.44	0.16
1,414	1.5	1,500	−0.21	0.048	1.31	1.84	0.031	5.95	1.41	0.13
σ				0.004	0.14	0.34	0.009	0.80	0.16	0.02
1,415	1.5	1,500	−0.39	0.454	1.18	6.99	0.040	0.88	1.04	0.21
σ				0.028	0.13	0.66	0.005	0.11	0.14	0.05
1,416	1.5	1,500	0.88	0.006	1.39	2.70	0.067	4.92	1.52	0.23
σ				0.0005	0.18	0.25	0.007	0.56	0.19	0.02
1,417	1.5	1,500	1.06	0.007	1.20	3.11	0.150	2.71	1.19	0.50
σ				0.001	0.24	0.52	0.042	0.48	0.21	0.30

Partition coefficients in weight ratio. σ is calculated from error propagation.
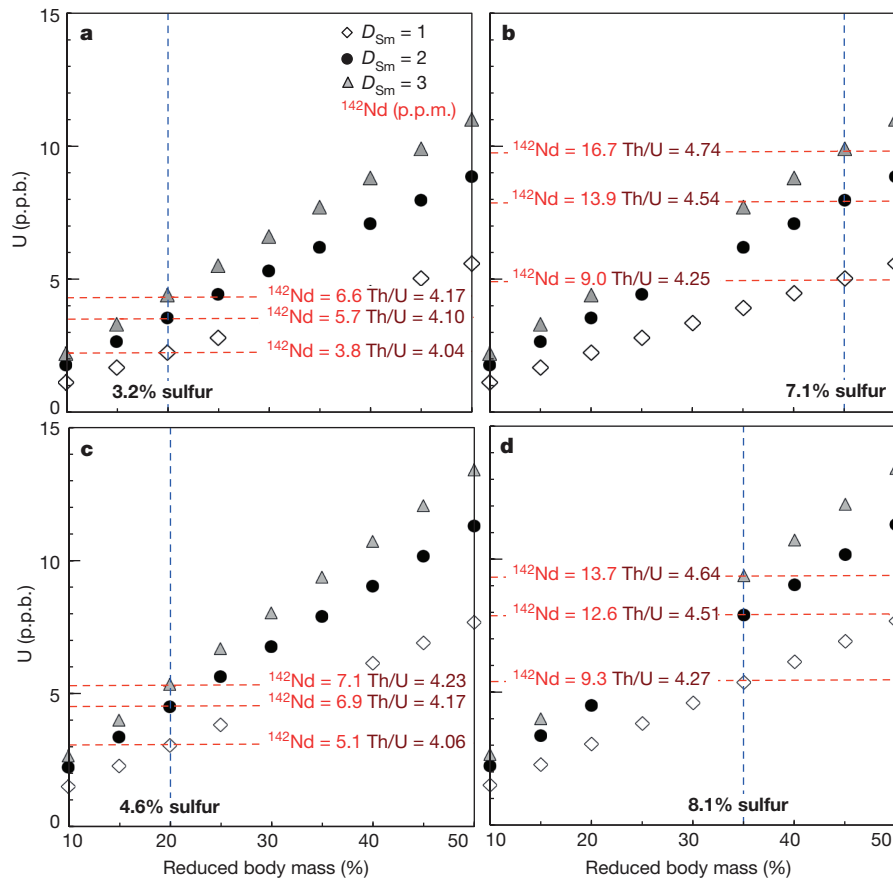
Figure 2 | Core content of U (p.p.b.) and mantle ^{142}Nd anomaly (p.p.m.). a, Calculated effect of adding to the growing Earth a reduced body of 20% of Earth's mass containing 0.15 mass fraction sulfide. The sulfide is added to the core and the silicate to the mantle. Sulphide−silicate D_U/D_{Sm} is fixed at 2, D_{Nd}/D_{Sm} at 1.4 and D_{Th}/D_U at 0.1 (Table 1). **b**, Same as **a** except the mass of reduced body is 45% of the Earth's mass. **c** and **d**, As for **a** and **b** except the reduced body contains 0.22 mass fraction sulfide. The reduced body and remainder of Earth each contain 14 p.p.b. U and 53.5 p.p.b. Th, consistent with chondritic abundances. Sulfide extraction was assumed to take place at the origin of the Solar System.

close to those summarized above if the Th/U of the bulk silicate Earth is constrained to be ≤4.5 or ≤4.7.

We note that the scenarios shown in Fig. 2 refer to a terrestrial core containing between 3.2 wt% S and 8.1 wt% S. The concentration of cosmochemically abundant volatile S in the core is unknown, but recent suggestions range from a cosmochemical estimate of 1.7 wt% (ref. 14) to ~6 wt% (ref. 28) from liquid-metal density measurements and 14.7 wt% (ref. 29) from high-temperature, high-pressure equation-of-state measurements. The range shown in Fig. 2 is, therefore appropriate for the current state of knowledge.

We conclude that a period of growth of the accreting Earth under reduced, S-rich conditions would generate a measureable (~+14 p.p.m.) ^{142}Nd anomaly in the silicate Earth, in agreement with observations. This would also add sufficient U and Th to the core to generate 2−2.4 TW of the energy required to drive the geodynamo.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Boyet, M. & Carlson, R. W. ^{142}Nd evidence for early (>4.53 Ga) global differentiation of the silicate Earth. *Science* **309**, 576−581 (2005).
2. Boyet, M. & Carlson, R. W. A new geochemical model for the Earth's mantle inferred from (SM)-S-146-Nd-142 systematics. *Earth Planet. Sci. Lett.* **250**, 254−268 (2006).
3. Campbell, I. H. & O'Neill, H. S. C. Evidence against a chondritic Earth. *Nature* **483**, 553−558 (2012).
4. Wade, J. & Wood, B. J. The Earth's 'missing' niobium may be in the core. *Nature* **409**, 75−78 (2001).
5. Kinoshita, N. *et al.* A shorter Sm-146 half-life measured and implications for Sm-146-Nd-142 chronology in the Solar System. *Science* **335**, 1614−1617 (2012).
6. Sprung, P., Kleine, T. & Scherer, E. E. Isotopic evidence for chondritic Lu/Hf and Sm/Nd of the Moon. *Earth Planet. Sci. Lett.* **380**, 77−87 (2013).
7. Blundy, J. D. & Wood, B. J. Prediction of crystal−melt partition coefficients from elastic moduli. *Nature* **372**, 452−454 (1994).
8. Palme, H. & O'Neill, H. S. C. in *The Mantle and Core Vol. 2 Treatise on Geochemistry* (ed. Carlson, R. W.) 1−38 (Elsevier, 2003).
9. Stein, C. A. in *Global Earth Physics: A Handbook of Physical Constants* (ed. Ahrens, T. J.) 144−158 (American Geophysical Union, 1995).
10. Labrosse, S., Poirier, J. P. & Le Mouel, J. L. The age of the inner core. *Earth Planet. Sci. Lett.* **190**, 111−123 (2001).
11. Nimmo, F., Price, G. D., Brodholt, J. & Gubbins, D. The influence of potassium on core and geodynamo evolution. *Geophys. J. Int.* **156**, 363−376 (2004).
12. Gessmann, C. K. & Wood, B. J. Potassium in the Earth's core? *Earth Planet. Sci. Lett.* **200**, 63−78 (2002).
13. Murthy, V. M., van Westrenen, W. & Fei, Y. W. Experimental evidence that potassium is a substantial radioactive heat source in planetary cores. *Nature* **423**, 163−165 (2003).
14. Dreibus, G. & Palme, H. Cosmochemical constraints on the sulfur content in the Earth's core. *Geochim. Cosmochim. Acta* **60**, 1125−1130 (1996).
15. Murrell, M. T. Partitioning of K, U, and Th between sulfide and silicate liquids— implications for radioactive heating of planetary cores. *J. Geophys. Res.* **91**, 8126−8136 (1986).
16. Wheeler, K. T., Walker, D., Fei, Y. W., Minarik, W. G. & McDonough, W. F. Experimental partitioning of uranium between liquid iron sulfide and liquid silicate: implications for radioactivity in the Earth's core. *Geochim. Cosmochim. Acta* **70**, 1537−1547 (2006).
17. Kiseeva, E. S. & Wood, B. J. A simple model for chalcophile element partitioning between sulphide and silicate liquids with geochemical applications. *Earth Planet. Sci. Lett.* **383**, 68−81 (2013).
18. Wade, J. & Wood, B. J. Core formation and the oxidation state of the Earth. *Earth Planet. Sci. Lett.* **236**, 78−95 (2005).
19. Rubie, D. C. *et al.* Heterogeneous accretion, composition and core-mantle differentiation of the Earth. *Earth Planet. Sci. Lett.* **301**, 31−42 (2011).
20. Nittler, L. R. *et al.* The major-element composition of Mercury's surface from MESSENGER X-ray spectrometry. *Science* **333**, 1847−1850 (2011).
21. Smith, D. E. *et al.* Gravity field and internal structure of Mercury from MESSENGER. *Science* **336**, 214−217 (2012).
22. Gannoun, A., Boyet, M., El Goresy, A. & Devouard, B. REE and actinide microdistribution in Sahara 97072 and ALHA77295 EH3 chondrites: a combined cosmochemical and petrologic investigation. *Geochim. Cosmochim. Acta* **75**, 3269−3289 (2011).
23. Blichert-Toft, J., Zanda, B., Ebel, D. S. & Albarede, F. The Solar System primordial lead. *Earth Planet. Sci. Lett.* **300**, 152−163 (2010).
24. Hofmann, A. W. Reviving the layered mantle: plan D. *American Geophysical Union Fall Meeting* Abstract V43E−05 http://abstractsearch.agu.org/meetings/2011/FM/V43E-05.html (2011).
25. Allègre, C. J., Dupre, B. & Lewin, E. Thorium uranium ratio of the Earth. *Chem. Geol.* **56**, 219−227 (1986).
26. Tatsumoto, M. Isotopic composition of lead in oceanic basalt and its implication to mantle evolution. *Earth Planet. Sci. Lett.* **38**, 63−87 (1978).
27. Elliott, T., Zindler, A. & Bourdon, B. Exploring the kappa conundrum: the role of recycling in the lead isotope evolution of the mantle. *Earth Planet. Sci. Lett.* **169**, 129−145 (1999).
28. Morard, G. *et al.* The Earth's core composition from high pressure density measurements of liquid iron alloys. *Earth Planet. Sci. Lett.* **373**, 169−178 (2013).
29. Seagle, C. T., Campbell, A. J., Heinz, D. L., Shen, G. Y. & Prakapenka, V. B. Thermal equation of state of Fe3S and implications for sulfur in Earth's core. *J. Geophys. Res.* **111**, B06209 (2006).

Author Contributions Both authors performed experiments and microanalysis, with about two-thirds done by A.W. Both authors contributed to writing the manuscript, with about two-thirds done by B.J.W.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.W. (anke.wohlers@earth.ox.ac.uk) or B.J.W. (berniew@earth.ox.ac.uk).

METHODS

Experimental methods. Starting materials for high-pressure experiments consisted of mixtures of ~50 wt% (Fe,Ni)S and ~50% of a synthetic silicate approximating the 1.5 GPa eutectic composition in the anorthite−diopside−forsterite system[30]. The sulfide component was analytical-grade FeS doped with 1%−3% NiS. Trace elements were added as a mix consisting of Zr, La, Ce, Nd, Sm, Eu, Yb, Th and U as oxides. After adding the trace-element mix such that each element was present at 1,000−2,000 p.p.m., the silicate and sulfide starting materials were mixed in 50:50 proportions and ground under acetone for 20 min, then dried at 110 °C before the experiment. Starting compositions were loaded into 3 mm outer diameter and 1 mm inner diameter graphite capsules.

Experiments were conducted in a half-inch-diameter piston-cylinder apparatus using external cylinders either of BaCO$_3$-silica glass (at 1,500 °C and 1,650 °C) or CaF$_2$ (at 1,400 °C) and an 8 mm outer diameter graphite furnace with a 1-mm-thick wall. The unsealed capsule was separated from the graphite furnace by an interior MgO sleeve, with a 0.5-mm-thick alumina disk on top to prevent puncture by the thermocouple. Temperatures were controlled and monitored using a tungsten−rhenium thermocouple (W5%Re/W26%Re), and the temperature was maintained within ±1 °C. Experimental conditions were 1,400 °C, 1,500 °C and 1,650 °C at 1.5 GPa and with experiment durations between 1 h and 4.5 h. These times are sufficient to approach equilibrium in small graphite capsules[17]. Experiments were quenched by turning off the power supply. After quenching, the capsule was extracted from the furnace, mounted in acrylic and polished for further analyses with electron microprobe and LA-ICP-MS. All experimental charges contained sulfide blebs embedded in a silicate glass matrix.

Microanalysis. Samples were analysed on the JEOL 8600 electron microprobe in the Archaeology Department at the University of Oxford. Wavelength dispersive analyses of the major-element compositions of silicate glasses and sulfides were performed at 15 kV with a beam current of 20 nA and a 10 µm defocused beam (Extended Data Tables 1 and 2). At least 20 analyses were taken of the silicate and sulfide in each experiment. Count times for major elements (Si, Al, Ca, Mg, Fe in silicate, Fe in sulfide) were 30 s on the peak and 15 s background. Minor elements (S, Ni, O) were analysed for 60 s peak and 30 s background. We have previously noted Ni loss from similar experiments[17] and the principal reason for adding Ni was to provide an additional check on LA-ICP-MS analyses of the trace elements of interest (see below). A range of natural and synthetic standards was used for calibration. Standards for silicate were wollastonite (Si, Ca), jadeite (Al), periclase (Mg) and haematite (Fe). Standards for sulfides were Ni metal (Ni), galena (S) and haematite (Fe, O). Oxygen in the sulfides was determined using the Kα peak and LDE crystal.

We determined U and Sm contents of three product sulfides as a further check on the LA-ICP-MS analyses. In this case we measured the Mα peak for U and the Lα peak for Sm using standards of UO$_2$ and SmPO$_4$ respectively and a PET crystal. Operating conditions were 15 kV, 40 nA and a 10 µm beam. The count time for U was 120 s on peak and 60 s background. Sm was analysed for 150 s on the peak and 75 s background.

Trace elements in silicates and sulfides were measured by LA-ICP-MS employing a NexION 300 quadropole mass spectrometer coupled to a New Wave Research UP213 Nd:YAG laser at the University of Oxford. A laser repetition rate of 10 Hz and spot size of 25−50 µm were used for silicate glasses and sulfides (Extended Data Tables 3 and 4) with an energy density of ~12 J cm^{-2}. Operating in time-resolved mode, we employed 20 s of background acquisition, followed by ablation for 60 s. Between analyses we employed a 60−90 s 'wash-out' time. The following masses were counted: ^{24}Mg, ^{27}Al, ^{29}Si, ^{57}Fe, ^{60}Ni, ^{43}Ca, ^{90}Zr, ^{139}La, ^{140}Ce, ^{142}Nd, ^{152}Sm, ^{153}Eu, ^{174}Yb, ^{232}Th, ^{239}U. Our external standard was NIST610 glass and we typically collected three spectra of this at the beginning and end of each sequence of 10−15 unknowns. The BCR-2G standard was used as a secondary standard to check the accuracy of the calibration. Ablation yields were corrected by referencing to the known concentrations of Si and Ca (silicate glass) and Fe (sulfides), which had been determined by microprobe. Data reduction was performed off-line using the Glitter 4.4.3 software package (http://www.glitter-gemoc.com/) which enabled us to identify occasional sulfide inclusions in the silicate analyses. Since the Fe content of the NIST610 standard is only 460 p.p.m., the background is high and the matrices are very different, so cross-checks on the sulfide analyses were required. Therefore, we measured Ni with the electron microprobe and LA-ICP-MS. In agreement with Kiseeva and Wood[17], we observed no systematic offset between electron microprobe and LA-ICP-MS analyses for Ni (Extended Data Tables 2 and 4). Additionally, as discussed above, the U and Sm contents of the sulfides were measured by electron microprobe in three experimental charges (numbers 1415, 464, and 477). Between 20 and 43 electron probe analyses were performed on each sample. The highest U and Sm concentrations were measured in experiment 1415 with LA-ICP-MS (U = 2,958 p.p.m., Sm = 719 p.p.m.). Comparative measurements with electron probe yielded values of U = 3,280 ± 490 p.p.m. and Sm = 707 ± 110 p.p.m., (uncertainty is 2 s.e.) and therefore show excellent agreement. Two samples with lower U and Sm concentrations were also analysed. LA-ICP-MS measurements for experiment 464 yielded U = 952 p.p.m. and Sm = 327 p.p.m., while experiment 477 gave U = 927 p.p.m. and Sm = 300 p.p.m. Electron microprobe concentrations of U = 1,164 ± 224 p.p.m. and Sm = 319 ± 87 p.p.m. (experiment 464) and U = 991 ± 69 p.p.m. and Sm = 277 ± 38 p.p.m. (experiment 477) are also in excellent agreement with the LA-ICP-MS measurements. We conclude that our LA-ICP-MS results have no detectable systematic offset due to matrix effects or calibration errors.

30. Presnall, D. C. *et al.* Liquidus phase relations on the join diopside-forsterite-anorthite from 1 atm to 20 kbar: their bearing on the generation and crystallization of basaltic magma. *Contrib. Mineral. Petrol.* **66,** 203−220 (1978).
31. Salters, V. J. M. & Stracke, A. Composition of the depleted mantle. *Geochem. Geophys. Geosyst.* **5,** http://dx.doi.org/10.1029/2003gc000597 (2004).

LETTER

Recovery potential of the world's coral reef fishes

M. Aaron MacNeil[1,2,3], Nicholas A. J. Graham[3], Joshua E. Cinner[3], Shaun K. Wilson[4,5], Ivor D. Williams[6], Joseph Maina[7,8], Steven Newman[9], Alan M. Friedlander[10,11], Stacy Jupiter[8], Nicholas V. C. Polunin[9] & Tim R. McClanahan[8]

Continuing degradation of coral reef ecosystems has generated substantial interest in how management can support reef resilience[1,2]. Fishing is the primary source of diminished reef function globally[3–5], leading to widespread calls for additional marine reserves to recover fish biomass and restore key ecosystem functions[6]. Yet there are no established baselines for determining when these conservation objectives have been met or whether alternative management strategies provide similar ecosystem benefits. Here we establish empirical conservation benchmarks and fish biomass recovery timelines against which coral reefs can be assessed and managed by studying the recovery potential of more than 800 coral reefs along an exploitation gradient. We show that resident reef fish biomass in the absence of fishing (B_0) averages ~1,000 kg ha^{-1}, and that the vast majority (83%) of fished reefs are missing more than half their expected biomass, with severe consequences for key ecosystem functions such as predation. Given protection from fishing, reef fish biomass has the potential to recover within 35 years on average and less than 60 years when heavily depleted. Notably, alternative fisheries restrictions are largely (64%) successful at maintaining biomass above 50% of B_0, sustaining key functions such as herbivory. Our results demonstrate that crucial ecosystem functions can be maintained through a range of fisheries restrictions, allowing coral reef managers to develop recovery plans that meet conservation and livelihood objectives in areas where marine reserves are not socially or politically feasible solutions.

There is widespread agreement that local and global drivers need to be addressed to reduce the degradation of coral reef ecosystems worldwide[1,2]. Numerous reef fisheries are so severely overexploited that critical ecosystem functions such as herbivory and predation are at risk[3–5]. Attempts to rebuild reef fish abundances and associated functions require clear timeframes over which assemblages can be restored, and viable management alternatives, such as marine reserves or gear restrictions, that promote recovery. Here we develop the first empirical estimate of coral reef fisheries recovery potential, compiling data from 832 coral reefs across 64 localities (countries and territories; Fig. 1a) to: (1) estimate a global unfished biomass (B_0) baseline—that is, the expected density of reef fish on unfished reefs (kg ha^{-1}); (2) quantify the rate of reef fish biomass recovery in well-enforced marine reserves using space-for-time substitution; (3) characterize the state of reef fish communities within fished and managed areas in terms of depletion against a B_0 baseline; (4) predict the time required to recover biomass and ecosystem functions across the localities studied; and (5) explore the potential returns in biomass and function using off-reserve management throughout the broader reefscape.

We used a Bayesian approach to estimate jointly B_0 as the recovery asymptote from well-enforced marine reserves (where fishing is effectively prohibited; Fig. 1b) and the average standing biomass of unfished remote areas more than 200 km from human settlements (Fig. 1c). We first used a space-for-time analysis of recovery in well-enforced marine reserves that varied in age and controlled for available factors known to influence observed fish biomass, including local net primary productivity, the percentage of hard coral cover, water depth, and reserve size[6] (Fig. 1b). We then modelled B_0 by linking this recovery data with prior information[4] on B_0 and biomass from remote reefs (Fig. 1c), an approach that explicitly assumes that marine reserves have the potential to recover to such levels in the absence of complicating factors, such as poaching or disturbance, and are of appropriate size[6]. Globally, expected B_0 for diurnally active, resident reef fish was 1,013 (963, 1469) kg ha^{-1} (posterior median (95% highest posterior density intervals)), with a biomass growth rate (r_0) of 0.054 (0.01, 0.11) from an estimated initial biomass in heavily fished reefs of 158 (43, 324) kg ha^{-1} (Fig. 1). The wide uncertainty in absolute B_0 reflected variability in average biomass among remote localities (from ~500 to 4,400 kg ha^{-1}; log-scale coefficient of variation = 0.08; geometric coefficient of variation = 0.61) as well as differences in productivity, hard coral cover, and atoll presence among reefs (Extended Data Fig. 1). We found no evidence of data provider bias (Extended Data Fig. 2) and model goodness-of-fit showed no evidence of lack of fit (Bayesian $P = 0.521$; Extended Data Fig. 3).

The status of reef fish assemblages on fished reefs against a B_0 baseline varied considerably by locality and whether there were management restrictions on fishing activities. Fished reefs (those that lacked management restrictions) spanned a wide range of exploitation states, from heavily degraded in the Caribbean and western Pacific, to high-biomass in the remote but inhabited Pitcairn and Easter Islands (Fig. 2a). Although previous studies have assessed how global reef fish yields relate to human population density[7], we characterize, for the first time, the state of fished reefs against an empirical baseline. Of concern was that more than a third of the fished reefs sampled had biomass below $0.25 B_0$, a point below which multiple negative ecosystem effects of overfishing have been shown to occur in the western Indian Ocean[7]. Only two localities, in Papua New Guinea and Guam, were at or near $0.1 B_0$, a fisheries reference point assumed to indicate collapse[8]. Reef fish assemblages fared far better when fishing activities were restricted in some way, including limitations on the species that could be caught, the gears that could be used, and controlled access rights (Fig. 2b). None of the localities with fisheries restrictions had average biomass levels below $0.25 B_0$ and 64% were above $0.5 B_0$, although some individual reefs within localities were below this level (Fig. 2b).

Despite extensive research into the benefits and planning of marine reserves, there is limited understanding of how long it takes reef fishes to recover once protected from fishing, limiting the ability of decision-makers to navigate management trade-offs. To estimate recovery times for fished and restricted reefs under hypothetical protection from fishing, we used the empirical recovery curve from marine reserves to back-calculate posterior virtual reserve ages (VA_i) for each locality, given their estimated level of fish biomass. We estimated the expected age of reserves at 90% recovery ($AR_{0.9}$) and subtracted the virtual reserve ages to calculate reef-specific expected recovery times ($TR_{0.9,i}$) under full closure (that is, $TR_{0.9,i} = AR_{0.9} - VA_i$). By sampling these quantities from the posteriors

[1]Australian Institute of Marine Science, PMB 3 Townsville MC, Townsville, Queensland 4810, Australia. [2]Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada. [3]Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia. [4]Department of Parks and Wildlife, Kensington, Perth, Western Australia 6151, Australia. [5]Oceans Institute, University of Western Australia, Crawley, Western Australia 6009, Australia. [6]Coral Reef Ecosystems Division, NOAA Pacific Islands Fisheries Science Center, Honolulu, Hawaii 96818, USA. [7]Australian Research Council Centre of Excellence for Environmental Decisions (CEED), University of Queensland, Brisbane, St Lucia, Queensland 4074, Australia. [8]Wildlife Conservation Society, Marine Programs, Bronx, New York 10460, USA. [9]School of Marine Science and Technology, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [10]Fisheries Ecology Research Lab, Department of Biology, University of Hawaii, Honolulu, Hawaii 96822, USA. [11]Pristine Seas-National Geographic, Washington DC 20036, USA.
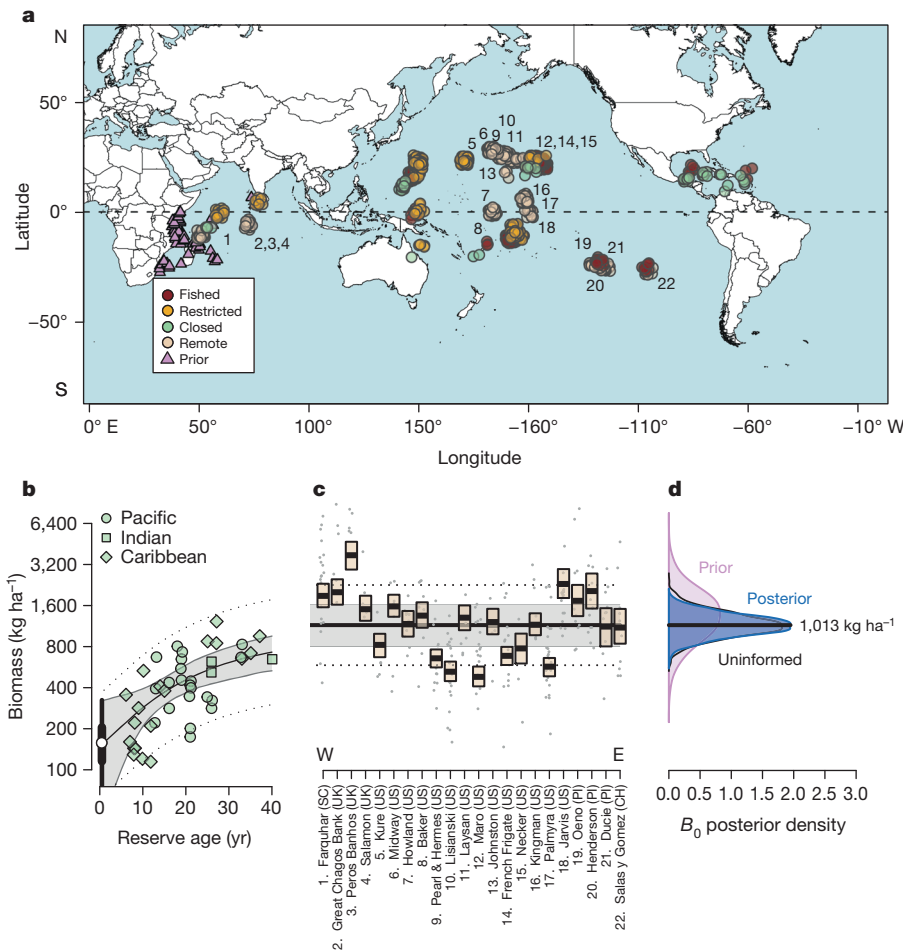
Figure 1 | Global reef fish biomass among management categories. a, Study ($n = 832$) and prior ($n = 157$) sites, with numbers matching graph in **c**. **b,** Posterior median recovery trajectory (black line) of reef fish biomass among reserve locations ($n = 45$), with 95% uncertainty intervals (grey), 95% prediction intervals (dotted line), estimated initial biomass (white circle with 50% (thick line) and 95% (thin line) highest posterior densities), and observed underwater visual census (UVC) data (green symbols). **c,** Posterior biomass for remote locations ($n = 22$; boxplots; 50% quantiles) with data (grey circles), median B_0 (black line), 95% uncertainty intervals (grey shading), and 95% prediction intervals (dotted line) from B_0 in **d**. **d,** Prior (violet), joint informed (dark blue), and uninformed (black line) posterior densities for B_0.

of our Bayesian model, we were able to develop probabilistic time frames for management along an expected path to recovery. Consistent with other studies on recovery benchmarks[9], and the United Nations Food and Agricultural Organization (FAO) definition of underexploited fisheries being between 0.8 and 1.0 (ref. 10), we defined recovered at 0.9 of B_0, but also estimated median recovery timeframes for a range of other recovery benchmarks and rates of increase (Methods).

On average, the fished and fishing-restricted reefs surveyed within localities would require 35 years of protection from fishing to recover to 0.9 B_0, while the most depleted reefs would require 59 years (Fig. 2c and Extended Data Fig. 4). Recovery times depended critically on the estimated rate of biomass recovery and the recovery benchmark used (Extended Data Fig. 5a). Although the influence of marine reserves can be detected within several years[11], our global analysis supports previous case studies[12,13] and a meta-analysis[14] showing comprehensive recovery of reef fish biomass probably takes decades to achieve. This suggests that most marine reserves, having been implemented in the past 10–20 years, will require many more years to achieve their recovery potential, underscoring the need for continued, effective protection and consideration of other viable management options.

To understand how the ecosystem functions provided by fishes change with protection from fishing, we examined relative changes in functional group biomass along the gradient from collapsed (101 (68, 144) kg ha^{-1}) to recovered (908 (614, 1,293) kg ha^{-1}), using generalized additive models to characterize trends. Despite substantial variability in the proportion of

each functional group among reefs, clear nonlinear trends were present in relative function (Extended Data Fig. 6). During initial recovery, functional returns of key low trophic level species increased rapidly, including browsers, scraper/excavators, grazers and planktivores (Fig. 2d and Extended Data Fig. 7). These are some of the most important ecosystem functions on coral reefs, as browsers and scraper/excavators promote coral dominance by controlling algae and clearing reef substrate for coral settlement and growth[15]; grazers help to limit the establishment of macroalgae by intense feeding on algal turfs[16]; and planktivores capture water-borne nutrients and sequester them to the reef food web[17]. Crucially, the relative functions of grazers and scrapers/excavators reached 80–100% of their maximum biomass by 0.5 B_0, while browsers, planktivores and the three top predator groups (macro-invertivores, pisci-invertivores and piscivores) increased steadily as standing biomass increased towards B_0. This overall pattern of functional change shows that key herbivore functions can be fulfilled at intermediate biomass levels, rather than solely among pristine areas.

Studies across gradients of human population and fishing densities have previously found the highest absolute losses of herbivores[5] and predators[18,19] can occur with relatively low fishing pressure; by contrast, our results show that the greatest functional changes occur when more than half of total biomass has been removed, supporting previous nonlinear relationships between biomass and function[4,16]. This disparity probably reflects differences in studying the effects of fishing on pristine versus altered reefs—where the apex predators not included in our analysis are readily
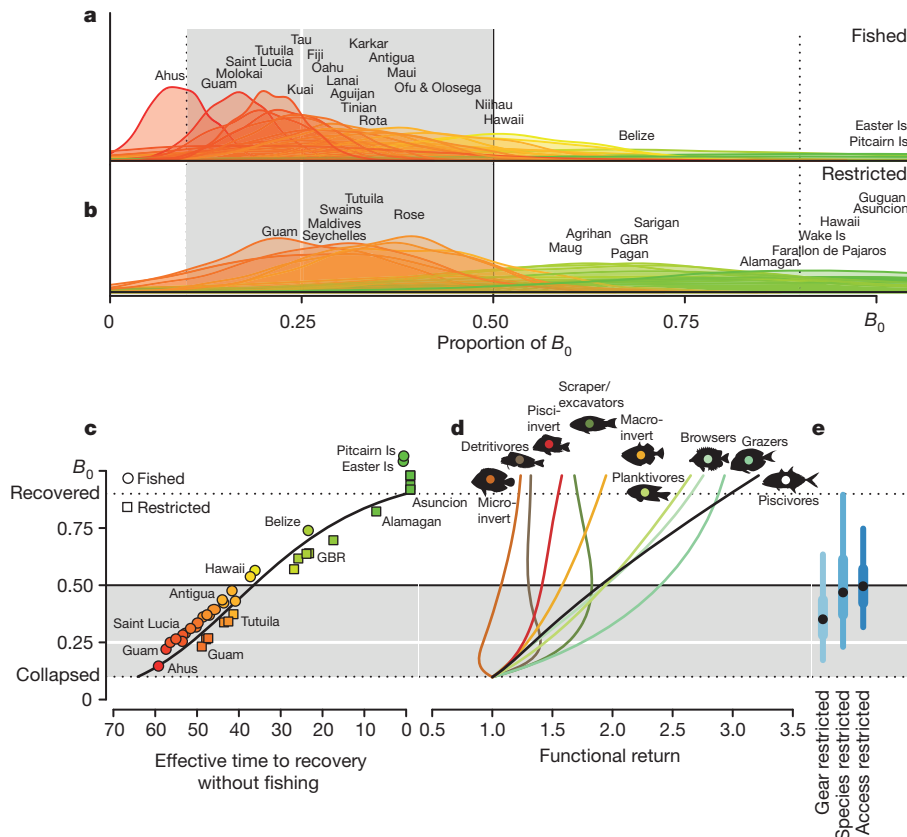
Figure 2 | Coral reef fish responses across the spectrum of potential recovery. a, b, Posterior density proportion of B_0 for fished ($n = 23$) (**a**) and fishing-restricted ($n = 17$) (**b**) coral reef locations, shaded from red (collapsed = $0.1\ B_0$) to green (recovered = $0.9\ B_0$). GBR, Great Britain; Is, islands. **c,** Expected times to recovery ($0.9\ B_0$) for fished (circles) and restricted (squares) reefs given full, effective closure. **d,** Average reef fish functional returns from collapsed to recovered. **e,** Median estimated proportion of B_0 among reef fishery management alternatives (black circles) with 50% (thick line) and 95% (thin line) uncertainty intervals.

removed[20]—and differences in socioeconomic conditions that influence reef exploitation at specific locations[21].

Although marine reserves have been widely advocated conservation tools[4], they can be untenable where people depend heavily on reef-based resources, highlighting the need for management alternatives to regulate fisheries on reefs. Therefore, to complement the use of effective marine reserves, we estimated expected biomass given alternative fishing restrictions (Fig. 2e), which typically receive less resistance from fishers than marine reserves[22]. On average, reefs with some form of fisheries restriction had biomass 27% higher than reefs open to fishing (Fig. 2a, b). Crucially, on reefs with bans on specific fishing gears, such as beach seines, or restrictions on the types of fish that can be caught, such as herbivores, biomass levels were between 0.3 and 0.4 B_0, the point at which up to 80% of herbivore function was retained (Fig. 2e). Thus, even simple fisheries restrictions can have substantial effects on fish functional groups that support important reef processes. Still greater biomass and functional returns were observed on reefs with access restrictions limiting the number of people allowed to fish a reef, such as family relations, or where other forms of established local marine tenure enable exclusion of external fishers[21]. Although these management alternatives clearly promote important functional gains relative to openly fished reefs, it is only among well-enforced, long-established marine reserves that predation is maximized, more than tripling the function of piscivory present on collapsed reefs.

The continuing degradation of the world's coral reefs underscores the need for tangible solutions that promote recovery and enhance ecosystem functions[4,23]. Our results demonstrate that well-enforced marine reserves can support a full suite of reef fish functions given enough time to recover. However, for reefs where marine reserves cannot be implemented, we find that ecosystem functions can be enhanced through various forms

of fisheries management. Addressing the coral reef crisis ultimately demands long-term, international action on global-scale issues such as ocean warming and acidification[24], factors that may diminish recovery potential by ~6% over the coming decades (Extended Data Fig. 5b). Despite these challenges, a range of fisheries management options is available to support reef resilience and it is likely that some combination of approaches will be necessary for success. Having benchmarks and timelines within an explicit biomass context, such as those provided here, increase the chances of agreeing on, and complying with, a mix of management strategies that will achieve conservation objectives while sustaining reef-based livelihoods.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Hughes, T. P. *et al.* Rising to the challenge of sustaining coral reef resilience. *Trends Ecol. Evol.* **25,** 633–642 (2010).
2. Graham, N. A. J. *et al.* Managing resilience to reverse phase shifts in coral reefs. *Front. Ecol. Environ* **11,** 541–548 (2013).
3. Dulvy, N. K., Freckleton, R. P. & Polunin, N. V. C. Coral reef cascades and the indirect effects of predator removal by exploitation. *Ecol. Lett.* **7,** 410–416 (2004).
4. McClanahan, T. R. *et al.* Critical thresholds and tangible targets for ecosystem-based management of coral reefs. *Proc. Natl Acad. Sci. USA* **108,** 17230–17233 (2011).
5. Bellwood, D. R. *et al.* Human activity selectively impacts the ecosystem roles of parrotfishes on coral reefs. *Proc. R. Soc. Lond. B* **279,** 1621–1629 (2012).
6. Edgar, G. J. *et al.* Global conservation outcomes depend on marine protected areas with five key features. *Nature* **506,** 216–220 (2014).

7. Newton, K. *et al.* Current and future sustainability of island coral reef fisheries. *Curr. Biol.* **17,** 655–658 (2007).
8. Worm, B. *et al.* Rebuilding global fisheries. *Science* **325,** 578–585 (2009).
9. Lambert, G. I. *et al.* Quantifying recovery rates and resilience of seabed habitats impacted by bottom fishing. *J. Appl. Ecol.* **51,** 1326–1336 (2014).
10. Worm, B. & Branch, T. A. The future of fish. *Trends Ecol. Evol.* **27,** 594–599 (2012).
11. Babcock, R. C. *et al.* Decadal trends in marine reserves reveal differential rates of change in direct and indirect effects. *Proc. Natl Acad. Sci. USA* **107,** 18256–18261 (2010).
12. McClanahan, T. R., Graham, N. A. J., Calnan, J. & MacNeil, M. A. Towards pristine biomass: reef fish recovery in coral reef marine protected areas in Kenya. *Ecol. Appl.* **17,** 1055–1067 (2007).
13. Russ, G. R. & Alcala, A. C. Decadal-scale rebuilding of predator biomass in Philippine marine reserves. *Oecologia* **163,** 1103–1106 (2010).
14. Molloy, P. P., McLean, I. B. & Côté, M. Effects of marine reserve age on fish populations: a global meta-analysis. *J. Anim. Ecol.* **46,** 743–751 (2009).
15. Rasher, D. B. *et al.* Consumer diversity interacts with prey defenses to drive ecosystem function. *Ecology* **94,** 1347–1358 (2013).
16. Mumby, P. J. *et al.* Empirical relationships among resilience indicators on Micronesian reefs. *Coral Reefs* **32,** 213–226 (2013).
17. Hamner, W. H. *et al.* Export-import dynamics of zooplankton on a coral reef in Palau. *Mar. Ecol. Prog. Ser.* **334,** 83–92 (2007).
18. Jennings, S. *et al.* Effects of fishing effort and catch rate upon the structure and biomass of Fijian reef fish communities. *J. Appl. Ecol.* **33,** 400–412 (1996).
19. Dulvy, N. K. *et al.* Size structural change in lightly exploited coral reef fish communities: evidence for weak indirect effects. *Can. J. Fish. Aquat. Sci.* **61,** 466–475 (2004).
20. Robbins, W. D., Hisano, M., Connolly, S. R. & Choat, J. H. Ongoing collapse of coral-reef shark populations. *Curr. Biol.* **16,** 2314–2319 (2006).
21. Cinner, J. E. *et al.* Co-management of coral reef social-ecological systems. *Proc. Natl Acad. Sci. USA* **109,** 5219–5222 (2012).
22. McClanahan, T. R., Abunge, C. A. & Cinner, J. E. Heterogeneity in fishers' and managers' preferences towards management restrictions and benefits in Kenya. *Environ. Conserv.* **39,** 357–369 (2012).
23. Bellwood, D. R., Hughes, T. P., Folke, C. & Nyström, M. Confronting the coral reef crisis. *Nature* **429,** 827–833 (2004).
24. Hoegh-Guldberg, O. *et al.* Coral reefs under rapid climate change and ocean acidification. *Science* **318,** 1737–1742 (2007).

Author Contributions M.A.M. conceived of the study with N.A.J.G., N.V.C.P., T.R.M., S.K.W. and J.E.C.; M.A.M. developed and implemented the analysis; M.A.M. led the manuscript with N.A.J.G., J.E.C. and S.K.W. All other authors contributed data and made substantive contributions to the text.

Author Information This is Social Ecological Research Frontiers (SERF) working group contribution number 10. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.M. (a.macneil@aims.gov.au).

METHODS

Reef fish biomass estimates were based on instantaneous visual counts (UVC) from 2,096 surveys collected from coral reef slopes (that is, the sloping, windward outer reef, selected specifically to standardize the reef habitat and remove potential bias associated with habitat type) on 832 individual reef sites (hereafter 'reef'). No statistical methods were used to predetermine sample size. All data were collected using standard belt-transects (50×5 m or 30×4 m) or point-counts (7 m radius) between 2002 and 2013, with the bulk of the data (92%) collected since 2006 (Supplementary Table 1). Data from belt transects and point counts have repeatedly been shown to be comparable in estimating fish abundance[25] and biomass[26]. Within each survey area, reef associated fishes were identified to species level, abundance counted, and total length estimated to the nearest 5 cm. A single experienced observer collected data for each data set except the NOAA data from the Pacific where multiple observers operate on every sampling mission. However, NOAA has extensive protocols in place to ensure that their observers are well trained and follow consistent protocols, ensuring the data are consistent and unbiased. We tested for any bias among data providers (capturing information on both inter-observer differences, and census methods) by including each data provider as a random effect in our model (see below), which assumes that there are inherent correlations within data sets that affect the means and associated errors estimated from their data. This analysis showed that there was no bias among data providers and that there is little information present in data provider identities (Extended Data Fig. 2). From these transect-level data, we retained counts of diurnally active, non-cryptic reef fish that are resident on the reef slope, excluding sharks and semi-pelagics (Supplementary Table 2). Metadata for the surveys are within the James Cook University research data repository, the Tropical Data Hub (https://eresearch.jcu.edu.au/tdh).

Total biomass of fishes on each transect was calculated using published length–weight relationships or those available on FishBase (http://fishbase.org). During this process, we removed 35 transects in which divers were mobbed by behaviourally aggregating species (for example, *Acanthurus coeruleus*; $n = 34$) or high biomass aggregating species (*Bolbometopon muricatum*; $n = 1$) that led to potentially unreliable estimates of standing biomass according to the data provider. This truncated data set was averaged to the reef level (that is, transects within the same section of continuous reef)[27] forming 832 distinct reefs that formed the basic data for our study. The data were sampled from key coral regions around the world; however, the coral triangle, Brazil, West Africa and the Red Sea/Arabian Sea regions are not represented. Fish species were assigned to functional groups based on trophic guilds and dietary information from the literature and FishBase. A key scale in our analysis was 'locality', defined as reef areas from 10s to 100s of kilometres that generally correspond to individual nations and map closely onto ranges of human influence[27], within which reefs were nested for analysis. In this way our analysis consisted of three spatial scales: reef, locality and global.

We used the PyMC package[28] for the Python programming language to conduct our analysis, running the (Metropolis–Hastings) MCMC sampler for 10^6 iterations, with a 900,000 iteration burn in and a thinning rate of 100, leaving 1,000 samples in the posterior of each parameter; these long (relative to Gibbs sampling, for example) burn-in times are often required with a Metropolis–Hastings algorithm. Convergence was monitored by examining posterior chains and distributions for stability and by running five chains from different starting points and checking for convergence using Gelman–Rubin statistics[29] for parameters across multiple chains, all of which were at or close to 1, indicating good convergence of parameters across multiple chains.

We used multiple data sources, including remote areas, asymptotes of well enforced marine reserves, and prior information, to estimate unfished biomass (B_0) and time for recovery. Remote areas, defined as having no recent history of fishing and being more than 200 km from human settlement, informed local B_{0l} and global B_0, given reef-specific covariates x_{nj} thought to influence standing biomass that were available at most localities. These covariates included local net primary production (NPP)[30], average proportion of hard coral cover[31], depth of survey (m)[32], and having been collected on an atoll (0/1 dummy variable)[33]. NPP was calculated as ensemble mean of estimates based on two NPP algorithms applied on MODIS and SeaWIFS data (that is, Carbon-based Production Model-2 (CbPM2)[34] and Vertically Generalized Production Model (VGPM; http://orca.science.oregonstate.edu)[35]; mg C m^{-2} day^{-1}. Each of these reef-specific nuisance parameters were mean centred to offset the reef level observations relative to the main focus of our model—the B_{0l} estimates.

To ensure an appropriate sub-model structure was used, we evaluated fits of three potential linear and nonlinear relationships (linear, second-order polynomial, and third-order polynomial) for each continuous nuisance parameter. We selected the best-fitting relationship for each nuisance parameter individually based on having the lowest deviance information criteria (DIC) value (Extended Data Table 1) and then compared DIC values of a candidate model set having all combinations of each nuisance parameter to select a final model (Extended Data Table 2). We also examined the posterior residuals for each nuisance parameter sub-model to ensure no heteroscedasticity was present and that errors were normally distributed (Extended Data Fig. 8).

To recognize potential data provider methodological effects, we incorporated data-provider status in our B_0 estimates by adding a random effect ρ_j for data provider j in our Bayesian hierarchical model. These factors were included in a log-normal hierarchical model for B_0, given reef-scale observations $y_{il,r}$:

$$y_{il,r} \sim N\left(\mu_{il,r}, \sigma_l\right) \tag{1}$$

$$\mu_{il,r} = B_{0l} + \beta_1 x_{coral,i} + \beta_2 x_{coral,i}^2 + \beta_3 x_{coral,i}^3 + \beta_4 x_{atoll,i} + \beta_5 x_{production,i} \\ + \beta_6 x_{production,i}^2 + \beta_7 x_{production,i}^3 + \rho_j \tag{2}$$

$$B_{0l} \sim N(B_0, \sigma_b), \tag{3}$$

and weakly-informative priors

$$\beta_{1,\ldots,7} \sim N(0.0, 100) \tag{4}$$

$$\sigma_{l,b} \sim U(0.0, 100) \tag{5}$$

$$\rho_j \sim N(0.0, 100). \tag{6}$$

Because this study built on previous research conducted in the western Indian Ocean[7] we used the posterior B_0 estimate from that study as the prior for our analysis:

$$B_0 \sim LN(7.08, 0.46) \tag{7}$$

allowing us to build on existing knowledge by directly integrating information between studies. As a check for those averse to building on previous research in this way, we also ran the full model using an uninformative B_0 prior, resulting in highly similar inferences, albeit with marginally greater uncertainty than the informed estimates (6.92 (6.52, 7.27) log(kg ha^{-1}) informed; 6.82 (6.45, 7.23) log(kg ha^{-1}) uninformed), demonstrating that the observed data dominated the prior in our analysis.

To estimate times to biomass recovery, we relied on data from well-enforced, previously fished marine reserves from around the world (Fig. 1a) and used a space-for-time substitution approach, assuming the relationship between reserve age and standing biomass follows a standard logistic regression model and the same reef-scale offset terms as above:

$$y_{i,a} \sim N(\mu_a, \sigma_m) \tag{8}$$

$$\mu_a = \frac{B_0}{1 + (B_0 = \mu_0)/\mu_0} e^{-ra} + \beta_1 x_{coral,i} + \beta_2 x_{coral,i}^2 + \beta_3 x_{coral,i}^3 + \beta_4 x_{atoll,i} \\ + \beta_5 x_{production,i} + \beta_6 x_{production,i}^2 + \beta_7 x_{production,i}^3 + \rho_j \tag{9}$$

Here a is the age of the marine reserve in years; μ_0 is the average initial reserve biomass; and r the average rate of biomass increase. This model is less hierarchically explicit than equation (2) owing to the scarcity of global marine reserve biomass data, and relies on the key assumption that average reserve potential recovery is consistent, absent the reef-scale effects in the model. Notably, B_0 is the same as in equation (3) and the linear offsets $\beta_{1,\ldots,7}$ the same as in (2), meaning their effects were jointly estimated from both remote and marine reserve data. Therefore, B_0 is estimated from both the trajectory of marine reserves through time and from the average biomass of all areas defined a priori as being remote: B_0 is the asymptote in the reserve component of the model and the global mean in the remote component of the model. μ_0, the minimum biomass at reserve age zero, was given an uninformative $\sim U(1,10)$ prior that spanned the range of the data; the standard deviation σ_m was as in (5); $x_{size,i}$ was set to allow for potential effects of reserve size, thought to be an important component of reserve success[6].

Next we estimated standing reef fish biomass across a range of fished locations, again hierarchically, given observer effects and reef-level observations within each location:

$$y_{il,f} \sim N\left(\mu_{il,f}, \sigma_f\right) \tag{10}$$

$$\mu_{il,f} = B_{l,f} + \beta_1 x_{coral,i} + \beta_2 x_{coral,i}^2 + \beta_3 x_{coral,i}^3 + \beta_4 x_{atoll,i} + \beta_5 x_{production,i} \\ + \beta_6 x_{production,i}^2 + \beta_7 x_{production,i}^3 + \rho_j \tag{11}$$

$$B_{l,f} \sim N(0.0, 100) \tag{12}$$

Here the $B_{l,f}$ terms denote independent log-biomass priors per location as we did not assume any parent (hierarchical) structure among locations other than potential data-provider effects; the standard deviation prior for σ_f was as in (5). Note that fishing pressure is a continuous variable that implicitly underlies the observed differences in exploitation state outside of the factors included in our analysis.

To estimate the standing biomass across a range of management categories, z, we applied similar methods:

$$y_{il,z} \sim N\left(\mu_{il,z}, \sigma_z\right) \tag{13}$$

$$\mu_{il,z} = B_{l,z} + \beta_1 x_{\text{coral},i} + \beta_2 x_{\text{coral},i}^2 + \beta_3 x_{\text{coral},i}^3 + \beta_4 x_{\text{atoll},i} + \beta_5 x_{\text{production},i}$$
$$+ \beta_6 x_{\text{production},i}^2 + \beta_7 x_{\text{production},i}^3 + \rho_j \tag{14}$$

$$B_{l,z} \sim N(0.0,100) \tag{15}$$

As for the fished locations, the $B_{l,z}$ terms denote independent log-biomass priors per location and the standard deviation prior for σ_z was as in (5). Management alternative effects were calculated as the average of the location-level posteriors for each group. Note that some locations in the data (Agrihan, Alamagan, Asuncion, Farallon de Pajaros, Guguan, Maug, Pagan, Rose and Sarigan) were passively fishery-restricted owing to isolation limiting effort that could be directed at the resource and, as a trait that cannot be actively managed, we excluded these locations from this section of our analysis.

Overall model fit. We conducted posterior predictive checks for goodness of fit using Bayesian P values[36], whereby fit was assessed by the discrepancy between observed or simulated data and their expected values. To do this we simulated new data (y_i^{new}) by sampling from the joint posterior of our model (θ) and calculated the Freeman–Tukey measure of discrepancy for the observed (y_i^{obs}) or simulated data, given their expected values (μ_i):

$$D(\text{y}|\theta) = \sum_i \left(\sqrt{y_i} - \sqrt{\mu_i}\right)^2 \tag{16}$$

yielding two arrays of median discrepancies $D(y^{\text{obs}}|\theta)$ and $D(y^{\text{new}}|\theta)$ that were then used to calculate a Bayesian P value for our model by recording the proportion of times $D(y^{\text{obs}}|\theta)$ was greater than $D(y^{\text{new}}|\theta)$ (Extended Data Fig. 3). For models not showing evidence of being inconsistent with the observed data, $D(y^{\text{obs}}|\theta)$ will be greater than $D(y^{\text{new}}|\theta)$ 50% of the time, giving $P = 0.5$; for models showing evidence of being inconsistent with the observed data, $D(y^{\text{obs}}|\theta)$ will, by specification, be greater than (or less than) $D(y^{\text{new}}|\theta)$ 95% of the time.

Times to recovery. We capitalized on our integrated Bayesian model to estimate location-specific recovery times for fished and fishery-restricted reefs within the Bayesian MCMC scheme. First we calculated the average reserve age at recovery (that is, $0.9B_0$: $B_{0.9}$), given the posterior biomass rate of growth r and initial biomass of μ_0 (see posterior parameter estimates in Supplementary Table 3):

$$\text{AR}_{0.9} = \frac{\log\left[\left(\frac{B_0}{B_{0.9}} - 1\right)\bigg/\left(\frac{B_0 - \mu_0}{\mu_0}\right)\right]}{-r} \tag{17}$$

Next we calculated location-specific virtual reserve ages, given their estimated level of log-biomass:

$$\text{VA}_i = \frac{\log\left[\left(\frac{B_0}{B_{l,f/z}} - 1\right)\bigg/\left(\frac{B_0 - \mu_0}{\mu_0}\right)\right]}{-r} \tag{18}$$

and subtracted this from $\text{AR}_{0.9}$ to give an expected time to recovery for each location:

$$\text{TR}_{0.9,i} = \text{AR}_{0.9} - \text{VA}_i \tag{19}$$

Because these calculations were conducted within our MCMC scheme they included posterior uncertainties, given the data and our model.

Variable recovery targets. Our choice to define recovery at $0.9B_0$ was based on recent work on recovery in the North Sea[9] and being the midpoint at which individual fish stocks are considered underexploited by the United Nations Food and Agricultural Organization[10]. However, to explore how expected time to recovery was dependent on this choice and the estimated rate of biomass growth, we calculated average reserve ages at recovery ($\text{AR}_{x,y}$) using the median posterior B_0 and μ_0 values (in (17)) while systematically varying the proportion of B_0 defined as recovered (between 0.8 to 1.0) and the rate of biomass growth (between posterior 95% UI range of 0.012 and 0.11). The resulting surface plot showed exponential increases in reserve ages at recovery for slower biomass growth rates and higher values of defined recovery due to the asymptotic nature of the logistic growth model used. (Extended Data Fig. 5).

Potential effects of climate change on B_0. A key assumption of the conclusions drawn from our results is that factors affecting total potential B_0 will remain stable through time. Climate projections have been equivocal as to what might happen to tropical fisheries over the coming decades[37], primarily owing to uncertainty in how production[38] and hard coral habitat[39] is expected to change, as well as difficulty in modelling tropical coastal habitats[37]. Nonetheless, we used the estimated relationships of log-biomass to productivity and hard coral cover (Extended Data Fig. 1) to

explore changes in B_0 owing to declines in both environmental conditions, using the median posterior estimates from our Bayesian hierarchical model. Results showed that by 2040, given an expected 4% loss of primary productivity[38] and a 2% annual loss of coral cover[39], we would expect to see a 6% drop in B_0, to 953 kg ha^{-1} (Extended Data Fig. 5b).

Log versus arithmetic scales of estimation. By adopting a hierarchical approach we, in effect, chose to average over location-specific differences to make global-scale inferences. We elected to model our data on the log-scale, as per fisheries convention[40], because it normalized the variance around our hierarchical model, greatly improving the precision of model estimates and the convergence of our model fits.

A key related point in our analysis is that our posterior calculations for fractions of B_0 were all on the arithmetic scale, by exponentiating each location-scale estimate and dividing by e^{B0}. To see why this makes sense, taking the posterior estimates for log-biomass from Ahus, PNG (4.54) and B_0 (6.92), Ahus would have retained 4.54/6.92 = 0.66 unfished log-biomass but only $e^{4.54}/e^{6.92} = 0.09$ absolute biomass. Given that this is the most heavily exploited reef in our database and that fisheries conventions for defining collapsed and recovered are arithmetic, we retained the arithmetic for our posterior calculations.

Functional returns. To understand how relative reef fish function would be expected to vary over the recovery range from collapsed (101 (68, 143) kg ha^{-1}) through to recovery (908 (614, 1293) kg ha^{-1}), we modelled the average biomass of each functional group across this range (that is, log(101) to log(908) kg ha^{-1}) relative to their initial biomass values (that is, average biomass of each functional group at log(101) kg ha^{-1}). We deemed these relative changes in biomass 'functional returns' because they express relative increases in function that could be expected given log-scale increases in the total biomass of a given functional group on a coral reef. To do this, and allow for expected non-linarites in functional group responses (due to, for example, community interactions, resource dynamics, the shape of response to which is currently unknown for most functional groups) we fit a series of generalized additive models (GAMs) to the proportion of each functional group over the community recovery range (Extended Data Fig. 6) in models that included the same covariates as our Bayesian hierarchical model (NPP, average proportion of hard coral cover, depth of survey, and having been collected on an atoll). The form of the model was, for each functional group k:

$$y_{il,k} \sim N\left(\mu_{il,k}, \sigma_k\right) \tag{20}$$

$$\mu_{il,k} = \beta_{0l} + f_1\left(x_{\log-\text{biomass},i}\right) + \beta_1 x_{\text{coral},i} + \beta_2 x_{\text{atoll},i} + \beta_3 x_{\text{production},i} \tag{21}$$

$$\beta_{0l} \sim N(0.0,100). \tag{22}$$

with the smooth function $f_1\left(x_{\log-\text{biomass},i}\right)$ describing the nonlinear relationship between observed functional group proportions and total log-biomass. Dividing the fitted GAMs for each functional group by the proportion at collapse provided a measure of expected functional return for each group, where a functional return of 2.0 would mean there is twice the log-biomass of a given functional group present compared to initial conditions. The rationale for this approach was that, as our data span the full range from 0.1 to 0.9 B_0, we did not need to predict outside of the data, but rather uncover the potentially nonlinear changes in relative function for each group over this range. All GAMs were run using in the *GAMM* package in R (http://www.r-project.org), using default smooth parameters that provided consistent fits to a per 0.1 log-kg moving average.

Code availability. The data set used in this analysis can be obtained from the corresponding author on request, and combined with PyMC code in the Supplementary Methods to replicate our Bayesian hierarchical analysis.

25. Samoilys, M. A. & Carlos, G. Determining methods of underwater visual census for estimating the abundance of coral reef fishes. *Environ. Biol. Fishes* **57**, 289–304 (2000).
26. Watson, R. A. & Quinn, T. J. Performance of transect and point count underwater visual census methods. *Ecol. Modell.* **104**, 103–112 (1997).
27. MacNeil, M. A. & Connolly, S. R. in *Ecology of Fishes on Coral Reefs: The Functioning of and Ecosystem in a Changing World* Ch. 12 (ed. Mora, C.) (Cambridge Univ. Press, 2015).
28. Patil, A., Huard, D. & Fonnesbeck, C. J. PyMC: Bayesian stochastic modelling in Python. *J. Stat. Softw.* **35**, 1–81 (2010).
29. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
30. Chassot, E. *et al.* Global marine primary production constrains fisheries catches. *Ecol. Lett.* **13**, 495–505 (2010).
31. Graham, N. A. J. *et al.* Climate warming, marine protected areas and the ocean-scale integrity of coral reef ecosystems. *PLoS ONE* **3**, e3039 (2008).
32. Newman, M. J. H., Paredes, G. A., Sala, E. & Jackson, J. B. C. Structure of Caribbean coral reef communities across a large gradient of fish biomass. *Ecol. Lett.* **9**, 1216–1227 (2006).

33. Cinner, J. E., Graham, N. A. J., Huchery, C. & MacNeil, M. A. Global effects of local human population density and distance to markets on the condition of coral reef fisheries. *Conserv. Biol.* **27,** 453–458 (2013).
34. Westberry, T., Bherenfeld, M. J., Siegel, D. A. & Boss, E. Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochem. Cy.* **22,** GB2024 (2008).
35. Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **42,** 1–20 (1997).
36. Brooks, S. P., Catchpole, E. A. & Morgan, B. J. T. Bayesian animal survival estimation. *Stat. Sci.* **15,** 357–376 (2000).
37. Cheung, W. W. L. *et al.* Large-scale redistribution of maximum fisheries potential in the global ocean under climate change. *Glob. Change Biol.* **16,** 24–35 (2010).
38. Sarmiento, J. L. *et al.* Response of ocean ecosystems to climate warming. *Glob. Biogeochem. Cycles* **18,** GB3003 (2004).
39. Bruno, J. F. & Selig, E. R. Regional decline of coral cover in the indo-pacific: timing, extent, and subregional comparisons. *PLoS ONE* **3,** e711 (2007).
40. Walters, C. J. & Martell, S. J. D. *Fisheries Ecology and Management* (Princeton Univ. Press, 2004).

LETTER

Emotional learning selectively and retroactively strengthens memories for related events

Joseph E. Dunsmoor[1], Vishnu P. Murty[1], Lila Davachi[1] & Elizabeth A. Phelps[1,2]

Neurobiological models of long-term memory propose a mechanism by which initially weak memories are strengthened through subsequent activation that engages common neural pathways minutes to hours later[1]. This synaptic tag-and-capture model has been hypothesized to explain how inconsequential information is selectively consolidated following salient experiences. Behavioural evidence for tag-and-capture is provided by rodent studies in which weak early memories are strengthened by future behavioural training[2,3]. Whether a process of behavioural tagging occurs in humans to transform weak episodic memories into stable long-term memories is unknown. Here we show, in humans, that information is selectively consolidated if conceptually related information, putatively represented in a common neural substrate, is made salient through an emotional learning experience. Memory for neutral objects was selectively enhanced if other objects from the same category were paired with shock. Retroactive enhancements as a result of emotional learning were observed following a period of consolidation, but were not observed in an immediate memory test or for items strongly encoded before fear conditioning. These findings provide new evidence for a generalized retroactive memory enhancement, whereby inconsequential information can be retroactively credited as relevant, and therefore selectively remembered, if conceptually related information acquires salience in the future.

People are motivated to remember the episodic details of emotional events, because this information is useful for predicting and controlling important events in the future[4,5]. In contrast, there is often little motivation to remember insignificant details we accumulate throughout the day, since much of this information is not associated with anything particularly meaningful. We do not always know, however, when a meaningful event will occur. From an adaptive memory perspective it is therefore critical that seemingly inconsequential details be stored in memory, at least temporarily, in the event that this information acquires relevance some time later. In this way, initially weak memories can be strengthened if this information later gains meaning. However, since we rarely encounter the same exact stimuli in the same exact situations it is advantageous for memories of other closely related information, encoded before a meaningful event, to be remembered as well. Such a mechanism could explain how a highly emotional event enhances memory for a host of details encoded earlier that, at the time, did not appear to hold any significance. Here, we provide evidence of a generalized retroactive memory enhancement in humans that is selective to information conceptually related to a future emotional event.

For episodic details to persist in long-term memory requires memory stabilization through the process of consolidation. A neurobiological account of memory consolidation has proposed a synaptic tag-and-capture mechanism whereby new memories that are initially weak and unstable are tagged for later stabilization by long-term potentiation (LTP) processes[1]. This mechanism has been extended to the domain of hippocampus-dependent learning in rats to explain how weak behavioural training that would otherwise be forgotten will endure in memory following a new behavioural experience (for example, exposure to novelty)—an effect referred to as behavioural tagging[2,3,6,7].

Whether behavioural tagging occurs in human episodic memory is unknown. Evidence for such an effect would require that memory for older events that are related to subsequent experiences is selectively enhanced while other unrelated information encoded at the same time should not receive a retroactive memory benefit. While prior studies have shown post-encoding modulation of memory consolidation with increases in stress and arousal[8,9], these demonstrations do not provide evidence of specificity. Another strong test of this hypothesized process is to mitigate the potential for selective rehearsal by presenting information in the absence of any motivation or instruction to remember (incidental encoding) and conducting a surprise memory test. Finally, models of behavioural tagging predict memory strengthening for weak encoding, but not strong encoding[6,7,10,11]. Thus, a task designed to retroactively boost relatively weak episodic memories should not retroactively benefit memories that were already strongly encoded.

Taking these criteria into consideration, we investigated whether information is selectively remembered if conceptually related information is later made salient through an amygdala-dependent learning task; that is, a trial-unique form of Pavlovian fear conditioning[12,13]. The encoding session occurred in three phases on the same day (Fig. 1). In phase 1, subjects classified 60 distinct basic-level objects as animals or tools (30 each). Shock electrodes were not attached during phase 1 and there was no explicit motivation or instruction to remember any of the pictures. Shortly thereafter, in phase 2, electric shock electrodes were attached and 30 novel images from one category (conditioned stimulus or CS$^+$, animals or tools, counterbalanced) were paired with a shock (unconditioned stimulus) to the right wrist at a reinforcement rate of 66%, while 30 novel images from the other category (CS$^-$, tools or animals, respectively) were unpaired. Skin conductance responses were acquired during fear conditioning to evaluate discriminatory fear learning. After conditioning, in phase 3, electric shock electrodes were removed and subjects classified additional images of 30 animals and 30 tools. Surprise recognition memory tests were then administered after either a 24-h delay, a 6-h delay, or immediately after phase 3 (see Methods for additional experimental details). The use of separate object categories provides the ability to test for selective consolidation in a within-subjects design. That is, we can assess whether fear conditioning preferentially enhances long-term memory for items related to the CS$^+$ but encoded before the conditioning experience, before any knowledge that related information would acquire future salience.

Significant physiological evidence of fear conditioning in phase 2, as assessed with greater skin conductance responses to the CS$^+$ versus the CS$^-$ category exemplars, was observed in all groups (Extended Data Fig. 1 and Methods). Recognition memory was calculated using corrected recognition (number of hits minus the number of false alarms to the corresponding category). An ANOVA with CS (CS$^+$, CS$^-$) and phase (pre-conditioning, conditioning, post-conditioning) as repeated measures, and retrieval group (24 h, 6 h, immediate) as between-subjects factor, revealed a main effect of CS ($F_{1,86} = 18.82$, $P < 0.001$, $\eta_p^2 = 0.18$), phase ($F_{2,85} = 29.35$, $P < 0.001$, $\eta_p^2 = 0.36$), and group ($F_{2,86} = 11.82$, $P < 0.001$, $\eta_p^2 = 0.22$), as well as a significant phase × group ($F_{4,172} = 4.49$,

[1]Department of Psychology and Center for Neural Sciences, New York University, New York, New York 10003, USA. [2]Nathan Kline Institute, Orangeburg, New York 10962, USA.

Pre-conditioning   Pavlovian conditioning   Post-conditioning

or

Time
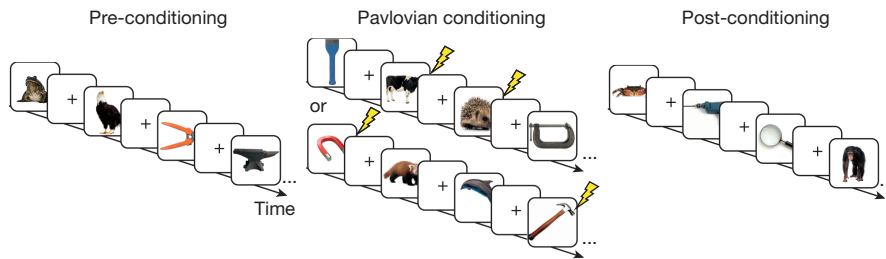
Figure 1 | Incidental encoding paradigm and example stimuli. Adult human subjects viewed 90 basic-level exemplars of animals and tools before, during and after fear conditioning. Before and after fear conditioning, subjects classified each object as an animal or a tool. During conditioning, electric shocks were paired with 20 out of 30 animal or tool pictures (counterbalanced between subjects) while subjects rated shock expectancy. A surprise recognition memory test was administered 24 h ($n = 30$), 6 h ($n = 30$), or immediately ($n = 29$) after encoding. Lightning bolts denote electric shocks.

$P = 0.002$, $\eta_p^2 = 0.09$) interaction. Follow-up planned ANOVAs and t-tests were conducted separately for the three retrieval groups.

The 24-h retrieval group (Fig. 2a and Extended Data Fig. 2) showed a main effect of CS ($F_{1,29} = 18.76$, $P < 0.001$, $\eta_p^2 = 0.39$) and phase ($F_{2,28} = 9.35$, $P = 0.001$, $\eta_p^2 = 0.40$). Follow up t-tests revealed that recognition memory was enhanced for CS^+ items encoded during fear conditioning ($t_{29} = 3.47$, $P = 0.002$, $d_{av} = 0.53$ (see Methods for an explanation of d_{av})), replicating previous findings[9]. This memory benefit extended to CS^+ exemplars encoded after fear conditioning, when the shock electrodes were unattached ($t_{29} = 3.42$, $P = 0.002$, $d_{av} = 0.58$), suggesting that selective effects of conditioning on subsequent memory can operate prospectively. Critically, a retroactive memory enhancement for CS^+ items was also observed. Memory was significantly stronger for items conceptually related to the CS^+ versus items related to the CS^- encoded before conditioning ($t_{29} = 2.48$, $P = 0.019$, $d_{av} = 0.41$), suggesting that weak memories from the pre-conditioning session were bolstered once conceptually related information acquired emotional relevance. There were no differences in false alarms between CS conditions ($P = 0.57$).

At 6-h retrieval (Fig. 2b), there was a main effect of CS ($F_{1,29} = 6.93$, $P = 0.01$, $\eta_p^2 = 0.19$) but no effect of phase ($P = 0.11$). The CS × phase interaction was significant ($F_{2,28} = 3.46$, $P = 0.05$, $\eta_p^2 = 0.19$). Follow-up t-tests showed significantly greater memory for CS^+ versus CS^- items encoded during pre-conditioning ($t_{29} = 2.41$, $P = 0.02$, $d_{av} = 0.40$) and fear conditioning ($t_{29} = 2.80$, $P = 0.009$, $d_{av} = 0.48$), replicating results obtained from 24-h retrieval. No differences between CS^+ and CS^- memory emerged at post-conditioning ($P = 0.52$), and there were no differences in false alarms between CS conditions ($P = 0.95$). This result indicates that fear-conditioning-mediated retroactive memory enhancements emerge by 6 h and are not dependent on sleep consolidation.

At immediate retrieval (Fig. 2c), there was no main effect of CS ($P = 0.17$), but there was a main effect of phase ($F_{2,27} = 20.32$, $P < 0.001$, $\eta_p^2 = 0.60$). Follow-up t-tests showed significantly greater memory for CS^+ versus CS^- items encoded during fear conditioning ($t_{28} = 2.14$, $P = 0.04$, $d_{av} = 0.50$). However, there was no difference in recognition memory between CS^+ and CS^- items encoded during pre-conditioning ($P = 0.97$) or post-conditioning ($P = 0.21$), and no differences in false alarms between CS conditions ($P = 0.74$). Importantly, this result suggests that fear-conditioning-mediated retroactive memory enhancement requires a period of consolidation. In order to directly assess whether the retroactive memory enhancement differed for delayed versus immediate retrieval, a memory difference score (corrected recognition for CS^+ items minus CS^- items) was calculated for all groups (Fig. 3). A comparison of delayed (24 and 6 h) versus immediate retrieval revealed significantly greater memory for CS^+ versus CS^- items encoded during pre-conditioning in the delayed groups relative to the immediate retrieval group ($t_{87} = 1.77$, $P = 0.04$, one-tailed, $d = 0.38$). In contrast to results from pre-conditioning, a comparison of post-conditioning memory between same-day (immediate and 6 h) versus next-day (24 h) retrieval revealed significantly greater memory for CS^+ versus CS^- items encoded during post-conditioning in the next-day group relative to the same-day groups ($t_{87} = 1.66$, $P = 0.05$, one-tailed, $d = 0.38$).

Models of behavioural tagging predict retroactive effects on weakly encoded memories, but no effect for strongly encoded memories[6,7,10,11]. To test whether strong encoding presents a boundary condition for retroactive enhancements of episodic memory, a separate group was
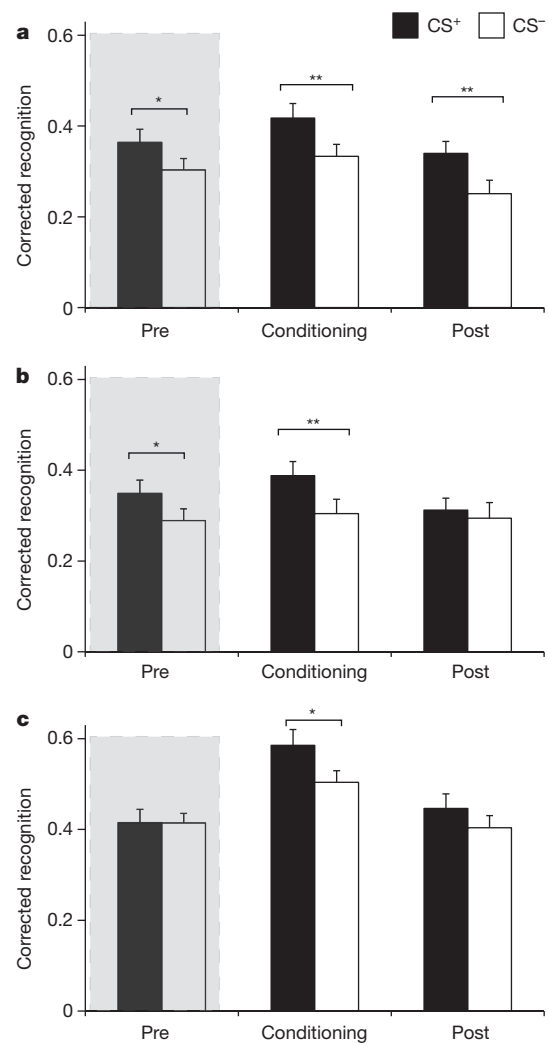
Figure 2 | Recognition memory performance. Memory at 24-h (**a**), 6-h (**b**), and immediate (**c**) retrieval showed enhanced corrected recognition memory for items from the CS^+ versus the CS^- category encoded during fear conditioning in all groups. However, memory was only retroactively enhanced for CS^+ items encoded during pre-conditioning following a 24-h or a 6-h delay. The shaded area highlights retroactive memory for items that preceded fear conditioning. CS^+, conditioned stimuli from the object category with exemplars paired with shock; CS^-, conditioned stimuli from the object category with exemplars never paired with shock. Error bars are s.e.m. *$P < 0.05$, **$P < 0.01$, two-tailed t-tests.
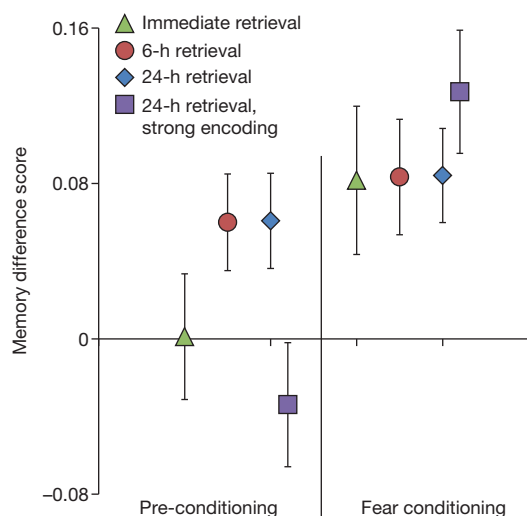
Figure 3 | Recognition memory difference scores. Corrected recognition difference scores (CS^+ minus CS^-) highlight that selective retroactive memory enhancements emerged at delay, but not immediate test, and not in subjects for whom pre-conditioning memory was strong before fear conditioning. Memory enhancements during fear conditioning were observed in all four experimental groups. Error bars are s.e.m.

shown each stimulus three times during pre-conditioning before fear conditioning. A surprise memory test was conducted 24 h later. An ANOVA showed a trend for CS ($P = 0.073$), an effect of phase ($F_{1,29} = 27.07, P < 0.001, \eta_p^2 = 0.49$), and a significant CS × phase interaction ($F_{1,29} = 17.04, P < 0.001, \eta_p^2 = 0.37$). Follow-up t-tests showed that this interaction was driven by greater corrected recognition memory for CS^+ (0.57 ± 0.03 (mean ± standard error)) than CS^- (0.45 ± 0.03) during fear conditioning ($t_{29} = 4.02, P = 0.02, d_{av} = 0.73$), and no difference between CS^+ (0.66 ± 0.04) and CS^- (0.69 ± 0.03) items encoded during pre-conditioning ($P = 0.29$) (Extended Data Fig. 3). A direct comparison between the 24-h weak- and 24-h strong-encoding groups revealed, as predicted, significantly greater overall (CS^+ and CS^-) memory for the strong-encoding (0.68 ± 0.03) versus weak-encoding (0.33 ± 0.02) group during pre-conditioning ($t_{58} = 3.87, P < 0.001, d = 0.99$), but a comparison of the memory difference score (CS^+ minus CS^-; Fig. 3) demonstrated a selective memory enhancement for CS^+ items in the 24-h weak-encoding group only ($t_{58} = 2.35, P = 0.01, d = 0.61$) (see Methods for additional analyses).

We found that memories for neutral information can be enhanced by a future emotional event that involves conceptually related material. The use of two category domains with relatively well-delineated neural substrates[14] allows us to speculate on a potential neurobiological mechanism mediating these effects. A recent neuroimaging investigation[13] showed that fear conditioning at the categorical level with animals and tools (akin to phase 2 from these experiments) modulates activity in category-selective regions in the extrastriate visual cortex; that is, activity in category-selective regions is enhanced in subjects for whom novel pictures of animals (or tools) predict shock. In the context of the present study, encoding during the pre-conditioning classification task may have set a weak learning tag in the hippocampus and these category-selective regions in the occipitotemporal cortex. Fear-conditioning-induced modulation of category-selective cortex and the hippocampus, via the amygdala or other regions involved in emotional learning circuitry, may then enhance related memories and possibly selectively prune unrelated memories[15]. Although these results are consistent with a putative tag-and-capture mechanism, whether such a mechanism explains the behavioural effect shown here requires future research. A consolidation mechanism is supported by the observation that memory enhancements for pre-conditioning were not seen in an immediate memory test. Notably, retroactive enhancements were evident after 6 h, in line

with studies showing that arousal-mediated consolidation effects are dependent on time, but not dependent on sleep[16,17]. This is in contrast to research showing selective retention for items retroactively made relevant through explicit instructions to remember, which finds effects only after a period of sleep consolidation[18,19].

This generalized retroactive memory enhancement can also be distinguished from prior studies of global post-encoding increases in consolidation through administration of stress or arousal[8,9], as emotional learning selectively enhanced memory for neutral items associated with that category, but not other neutral content encoded at the same time. By virtue of presenting information before Pavlovian conditioning, we can also disentangle enhanced attention at the time of encoding induced by the anticipation of shock from post-encoding consolidation processes. That is, during phase 1 there was no chance of receiving shocks (the shock leads were not attached), and no details had been provided to the subject about the contingencies of shock administration for later phases of the experiment (see Methods for further details). These results are also different from generalization that involves overlapping representations of cues pre-associated before reinforcement; for example, acquired equivalence[20]. In the present study, information presented at each phase of encoding is related at the conceptual level, but is never repeated or directly combined with information presented at another phase of incidental encoding.

Notably, while a retroactive memory benefit was shown after 24-h and 6-h delays, a proactive memory benefit was only demonstrated after 24 h. This finding was unexpected, and indicates that retroactive and proactive arousal-mediated memory enhancements are separable and perhaps rely on different mechanisms. In a potentially analogous finding[21], the ability to make inferential judgments regarding previously learned relational knowledge was reported to increase following a delay, and be further boosted following sleep. Whether the proactive memory enhancement in this study relies on a period of sleep consolidation is an intriguing possibility that may help dissociate mechanisms supporting retroactive versus proactive emotional memory effects.

In conclusion, our work provides new evidence for selective consolidation of information conceptually related to a future meaningful event. These findings support an implication proposed previously[1] in the formulation of the synaptic tag-and-capture mechanism, that late-phase LTP of synaptic activity could explain enhanced memories for seemingly insignificant details surrounding emotional events. An intriguing implication of this finding concerns the adaptive nature of episodic memory. Specifically, humans and other animals continuously monitor the environment, accumulating countless details. Much of this information is forgotten. However, meaningful events can selectively preserve memory for previously encountered information that seemed insignificant at the time it was encoded. Whether such a mechanism contributes to persistent intrusive memories and overgeneralization of fear characteristic of trauma and stress-related disorders merits further empirical research.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Frey, U. & Morris, R. G. Synaptic tagging and long-term potentiation. *Nature* **385**, 533–536 (1997).
2. Ballarini, F., Moncada, D., Martinez, M. C., Alen, N. & Viola, H. Behavioral tagging is a general mechanism of long-term memory formation. *Proc. Natl Acad. Sci. USA* **106**, 14599–14604 (2009).
3. de Carvalho Myskiw, J., Benetti, F. & Izquierdo, I. Behavioral tagging of extinction learning. *Proc. Natl Acad. Sci. USA* **110**, 1071–1076 (2013).
4. LaBar, K. S. & Cabeza, R. Cognitive neuroscience of emotional memory. *Nature Rev. Neurosci.* **7**, 54–64 (2006).
5. Lisman, J., Grace, A. A. & Duzel, E. A neoHebbian framework for episodic memory; role of dopamine-dependent late LTP. *Trends Neurosci.* **34**, 536–547 (2011).
6. Wang, S. H., Redondo, R. L. & Morris, R. G. Relevance of synaptic tagging and capture to the persistence of long-term potentiation and everyday spatial memory. *Proc. Natl Acad. Sci. USA* **107**, 19537–19542 (2010).

7. Moncada, D. & Viola, H. Induction of long-term memory by exposure to novelty requires protein synthesis: evidence for a behavioral tagging. *J. Neurosci.* **27,** 7476–7481 (2007).
8. McGaugh, J. L. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.* **27,** 1–28 (2004).
9. Cahill, L. & McGaugh, J. L. Mechanisms of emotional arousal and lasting declarative memory. *Trends Neurosci.* **21,** 294–299 (1998).
10. Moncada, D., Ballarini, F., Martinez, M. C., Frey, J. U. & Viola, H. Identification of transmitter systems and learning tag molecules involved in behavioral tagging during memory formation. *Proc. Natl Acad. Sci. USA* **108,** 12931–12936 (2011).
11. da Silva, B. M., Bast, T. & Morris, R. G. Spatial memory: behavioral determinants of persistence in the watermaze delayed matching-to-place task. *Learn. Mem.* **21,** 28–36 (2014).
12. Dunsmoor, J. E., Martin, A. & LaBar, K. S. Role of conceptual knowledge in learning and retention of conditioned fear. *Biol. Psychol.* **89,** 300–305 (2012).
13. Dunsmoor, J. E., Kragel, P. A., Martin, A. & LaBar, K. S. Aversive learning modulates cortical representations of object categories. *Cereb. Cortex* **24,** 2859–2872 (2014).
14. Martin, A. The representation of object concepts in the brain. *Annu. Rev. Psychol.* **58,** 25–45 (2007).
15. Stickgold, R. & Walker, M. P. Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neurosci.* **16,** 139–145 (2013).
16. Park, J. Effect of arousal and retention delay on memory: a meta-analysis. *Psychol. Rep.* **97,** 339–355 (2005).
17. Kleinsmith, L. J. & Kaplan, S. Paired-associate learning as a function of arousal and interpolated interval. *J. Exp. Psychol.* **65,** 190–193 (1963).
18. Wilhelm, I. *et al.* Sleep selectively enhances memory expected to be of future relevance. *J. Neurosci.* **31,** 1563–1569 (2011).
19. van Dongen, E. V., Thielen, J. W., Takashima, A., Barth, M. & Fernandez, G. Sleep supports selective retention of associative memories based on relevance for future utilization. *PLoS ONE* **7,** e43426 (2012).
20. Shohamy, D. & Wagner, A. D. Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* **60,** 378–389 (2008).
21. Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D. & Walker, M. P. Human relational memory requires time and sleep. *Proc. Natl Acad. Sci. USA* **104,** 7723–7728 (2007).

Author Contributions J.E.D. designed and conducted the study. J.E.D. and V.P.M. analysed the data. J.E.D., V.P.M., L.D. and E.A.P. interpreted the results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.D. (lila.davachi@nyu.edu) or E.A.P. (liz.phelps@nyu.edu).

METHODS

Participants. A total of 138 subjects were recruited to participate. Nineteen subjects were removed from the analysis for failure to return for the memory test ($n = 6$), failure to understand or follow the task instructions ($n = 7$), equipment failures with stimulus presentation software ($n = 3$), a failure to show any evidence of recognition memory above chance ($n = 2$), or indicating that the memory test was not a surprise ($n = 1$). The final sample included 119 subjects (Age = 23.42 ± 3.15 years (mean $\pm$ s.d.), 62 females). Subjects were assigned to 1 of 4 groups, immediate retrieval ($n = 29$, 16 females), 6-h retrieval ($N = 30$, 15 females), 24-h retrieval ($n = 30$, 20 females), or 24-h retrieval strong pre-conditioning encoding ($n = 30$, 11 females). Subjects in the immediate and 24-h retrieval groups were randomly assigned. The 6-h and 24-h strong-encoding groups were run as follow-up studies, and group assignment was not determined by randomization. Sample size was based on prior studies of categorical fear learning[12,13]. No statistical method was used to predetermine sample size. All subjects provided written informed consent approved by the University Committee on Activities Involving Human Subjects at New York University.

Behavioural paradigm and stimulus materials for 24-h, 6-h and immediate retrieval groups. The study involved two experimental sessions: incidental encoding and a surprise recognition memory test. The incidental-encoding session included 3 phases: pre-conditioning, fear conditioning, and post-conditioning. Each phase included 30 colour photographs of animals and 30 colour photographs of tools presented on a white background. Pictures were obtained from the website http://www.lifeonwhite.com or from publicly available resources on the internet. Each picture was a different basic-level exemplar with a different name; for example, there were not two different pictures of a dog. Stimulus order was counterbalanced across subjects and pseudo-randomized such that no more than 3 pictures from the same category appeared in a row.

During pre-conditioning, pictures were presented for 2.5 s with a 6 ± 2 s variable inter-trial interval that included a fixation cross on a blank background. The total duration of pre-conditioning was ~8.5 min. During pre-conditioning subjects made two-alternative forced-choice picture identifications ('animal' or 'tool'). Specifically, subjects were asked to classify each picture as either an animal or a tool by pressing the 1 or 2 button on a keypad on every trial. The buttons corresponding to animal and tool were counterbalanced across subjects.

Fear conditioning followed pre-conditioning ~5 min later. Between pre-conditioning and fear conditioning, shock leads were attached to the right wrist, and intensity was calibrated to a level deemed highly unpleasant, but not painful, using an ascending staircase procedure. Skin conductance response (SCR) leads were attached to the left palm. During fear conditioning, pictures were presented for 4.5 s with a variable inter-trial interval of 8 ± 2 s, which allowed time to measure SCRs before shocks occurred on CS^+ trials, and for SCRs to return to baseline after CS presentation. Shocks occurred on 20 out of 30 CS^+ trials at the end of the trial, co-terminating with the picture. The CS^+ trials paired with shock were counterbalanced between subjects. The total duration of fear conditioning was ~12 min. During fear conditioning, subjects made a two-alternative forced-choice shock expectancy rating (1 = shock, 2 = no shock). Specifically, subjects were asked to rate whether they expected the shock or not on every trial. Subjects were not instructed about the conditioned–unconditioned stimulus contingencies, and had to learn the category level association between the pictures and the shock through experience. Subjects were told explicitly that the button presses did not have any effect on whether or not the shock would occur, thus eliminating the chance for subjects to mistakenly attribute the outcome to their actions. The object categories serving as CS^+/CS^- were counterbalanced between subjects. After fear conditioning, the shock leads were removed and subjects were asked to rate the intensity of the shock on a scale from 1 (not at all unpleasant) to 10 (extremely unpleasant). The average rating was 6.17 (s.d. = 1.46), and there were no differences in mean intensity ratings between groups.

After fear conditioning the shock electrodes were removed. Post-conditioning occurred approximately 3 min after the end of fear conditioning. Procedures and instructions for post-conditioning were identical to those of pre-conditioning.

Recognition memory test procedures. The recognition memory test included the 90 CS^+ and 90 CS^- pictures seen the previous day, along with 90 new pictures of animals and 90 new pictures of tools (total of 360 pictures shown during the recognition memory test). The test was self-paced. Subjects rated whether each picture was new or old and their confidence by making 1 of 4 possible responses: definitely new, maybe new, maybe old, or definitely old. Memory responses were collapsed across confidence. Analysis focusing on high-confidence responses yielded similar results (all data are presented in Extended Data Tables 1–4). We performed our analysis on corrected recognition scores (hits minus false alarms) to account for differences in response criteria across participants. Data were normally distributed and variance was similar between groups. There were no differences in false alarms between CS categories (reported in main text).

Behavioural paradigm for strong encoding, 24-h retrieval. A separate group underwent a modified version of phase 1 encoding in which each stimulus ($n = 30$ animals, $n = 30$ tools) was presented 3 times each to strengthen memory for items encoded before fear conditioning. Trial order was randomized with the following constraints. First, no more than three images from the same object category appeared in a row. Second, each exemplar was presented twice during the first 120 trials, and once in the final 60 trials. Each picture was presented once during the final 60 trials to ensure that the lag between final stimulus presentation and conditioning was matched with the other protocols. To help ensure that the total duration of the experimental session was equivalent to the other groups, the inter-trial interval during phase 1 was reduced to 3.5 ± 0.5 s. Phase 1 was followed by the fear-conditioning protocol employed in the other groups (30 novel CS^+ and 30 novel CS^- trials, with 20/30 CS^+ trials paired with shock). As we were specifically interested in the effects on retroactive memory enhancements, we did not conduct a post-conditioning encoding session. This also helped keep the total time of the encoding session equivalent to the other experimental groups. The retrieval test for the strong-encoding group included the 60 CS^+ and 60 CS^- pictures seen the previous day, along with an equal number of new pictures from the CS^+ and CS^- categories (60 each).

Subject instructions and explicit knowledge regarding fear conditioning and memory test. Subjects were informed in advance that the study would involve electrical stimulation, and during informed consent each subject was told where the shock electrodes would be placed, and how the experimenter would calibrate the shock to a level they deemed highly unpleasant, but not painful. Importantly, no specific information was provided regarding the fear-conditioning phase before pre-conditioning, and shock leads were not attached during pre-conditioning. Consequently, even if subjects anticipated receiving shocks at a later phase of the experiment, this could not have a selective effect for one category of objects, since no details were provided regarding fear conditioning by this point of the task.

To assess whether subjects expected the surprise memory test, subjects in the 6-h retrieval group and the 24-h strong-encoding group were asked two questions when they returned for the memory test. First, they were asked, "Do you have any expectations of what this next task in the experiment will be: yes or no?" Subjects were then told that we would be conducting a test of their memory for the pictures they saw earlier, and were asked to indicate on a 5-point scale how surprised they were by a memory test, from 1 (I did not expect a memory test at all) to 5 (Yes, I knew there would be a memory test). The mean response was 2.63 (s.d. = 1.08). Only one subject responded "yes" to the first question and guessed correctly about a memory test. This was also the only subject to respond "5" on the second question. This subject was not included in the analysis.

Estimates of effect size. Effect sizes reported for ANOVAs in the manuscript are partial eta squared. For paired t-tests, we calculated Cohen's d using the mean difference score as the numerator and the average standard deviation of both repeated measures as the denominator, as suggested in ref. 22. This effect size is referred to in the text as d_{av}, where 'av' refers to the use of the average standard deviations in the calculation.

Shock and psychophysiology. A 200-ms shock was delivered to the right wrist using pre-gelled snap electrodes (BIOPAC EL508) connected to a Grass Medical Instruments stimulator (West Warwick, Rhode Island). SCR electrodes were placed on the hypothenar eminence of the palmar surface of the left hand using pre-gelled snap electrodes (BIOPAC EL509). Data were collected using a BIOPAC MP-100 System (Goleta, CA), and responses calculated using established criteria[23,24]. In brief, an SCR was considered related to CS presentation if the trough-to-peak deflection occurred 0.5–4.5 s following CS onset, lasted between 0.5 and 5.0 s, and was greater than 0.02 microsiemens (μS). Responses that did not fit these criteria were scored as zero. SCR values were obtained using a custom Matlab (The MathWorks, Inc.) script that extracted SCRs for each trial using the above criteria[25].

SCR results. SCRs were collected as a manipulation check that the fear-conditioning procedure effectively generated higher autonomic arousal on CS^+ trials than CS^- trials. SCR data was not analysed for 13 subjects due to equipment malfunction with the BIOPAC during data acquisition (24-h, $n = 4$; 6-h, $n = 3$; immediate, $n = 5$; 24-h strong, $n = 1$) and for 6 subjects due to an overall lack of measurable electrodermal responses (24-h, $n = 1$; 6-h, $n = 1$; immediate, $n = 2$; 24-h strong, $n = 2$). Paired t-tests showed enhanced SCRs to the CS^+ versus the CS^- in all four groups, and all P values were <0.0002, providing confirmation that the fear-conditioning manipulation was effective.

Supplementary memory analyses. To evaluate whether the memory enhancement observed for CS^+ versus CS^- items was different between the pre-conditioning and fear-conditioning phases, we compared the memory difference scores (CS^+ minus CS^-) between these two phases. This analysis was restricted to the two groups showing a selective CS^+ retroactive memory enhancement, the 24-h and 6-h delay groups. The memory difference score between pre-conditioning and fear conditioning was not different for either the 24-h ($P = 0.49$) or the 6-h ($P = 0.52$) delay

group. This analysis confirms that the CS^+ retroactive memory enhancement was not significantly different from the CS^+ fear-conditioning memory enhancement.

To ensure that the object categories serving as CS^+ and CS^- did not interact with memory effects, the object category subgroup (that is, animal CS^+/tool CS^-; tool CS^+/animal CS^-) was included as a covariate in a supplementary ANOVA. Subgroup did not interact with CS and phase for any group (all P values > 0.32).

As rodent studies of behavioural tagging show that the time interval between weak encoding and exposure to novelty can influence memory strength[3], we explored whether memory for items encoded during pre-conditioning (phase 1) were affected by the time relative to the start of fear conditioning (phase 2) in the 24-h retrieval group. For this analysis, items from pre-conditioning were binned according to tertiles corresponding to CS^+ (and CS^-) trials 0–10, 11–20 and 21–30. Tertiles roughly correspond to ~14 to 11, ~11 to 8, and ~8 to 5 min before the start of fear conditioning, respectively. An ANOVA on the CS^+ minus CS^- memory difference score using tertiles as a factor revealed a significant linear effect ($F_{1,29} = 4.40$,

$P = 0.044$, $\eta_p^2 = 0.13$), such that memory difference between CS^+ and CS^- trials diminished from the first tertile (11 ± 0.03 (mean ± s.e.m.)), to the second tertile (0.08 ± 0.04), to the third tertile (0.03 ± 0.03). This result suggests that the time between weak episodic encoding and emotional learning may influence the strength of retroactive memory enhancements.

22. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4,** 863 (2013).
23. Schiller, D. *et al.* Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463,** 49–53 (2010).
24. Dunsmoor, J. E., Mitroff, S. R. & LaBar, K. S. Generalization of conditioned fear along a dimension of increasing fear intensity. *Learn. Mem.* **16,** 460–469 (2009).
25. Green, S. R., Kragel, P. A., Fecteau, M. E. & LaBar, K. S. Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *Int. J. Psychophysiol.* **91,** 186–193 (2014).

Thirst driving and suppressing signals encoded by distinct neural populations in the brain

Yuki Oka[1,2]†, Mingyu Ye[1,2] & Charles S. Zuker[1,2]

Thirst is the basic instinct to drink water. Previously, it was shown that neurons in several circumventricular organs of the hypothalamus are activated by thirst-inducing conditions[1]. Here we identify two distinct, genetically separable neural populations in the subfornical organ that trigger or suppress thirst. We show that optogenetic activation of subfornical organ excitatory neurons, marked by the expression of the transcription factor ETV-1, evokes intense drinking behaviour, and does so even in fully water-satiated animals. The light-induced response is highly specific for water, immediate and strictly locked to the laser stimulus. In contrast, activation of a second population of subfornical organ neurons, marked by expression of the vesicular GABA transporter VGAT, drastically suppresses drinking, even in water-craving thirsty animals. These results reveal an innate brain circuit that can turn an animal's water-drinking behaviour on and off, and probably functions as a centre for thirst control in the mammalian brain.

Body fluid homeostasis regulates the internal salt and water balance; as this balance shifts, the brain senses these changes and triggers specific goal-oriented intake behaviours[2,3]. For instance, salt-deprived animals may actively consume salty solutions, even though such high levels of salt are normally strongly aversive[4–6]. Similarly, dehydrated animals are strongly motivated to consume water[7,8]. Previous studies have shown that various regions in the circumventricular organs (CVO) of the hypothalamus are activated in response to dehydration[1]. In addition,
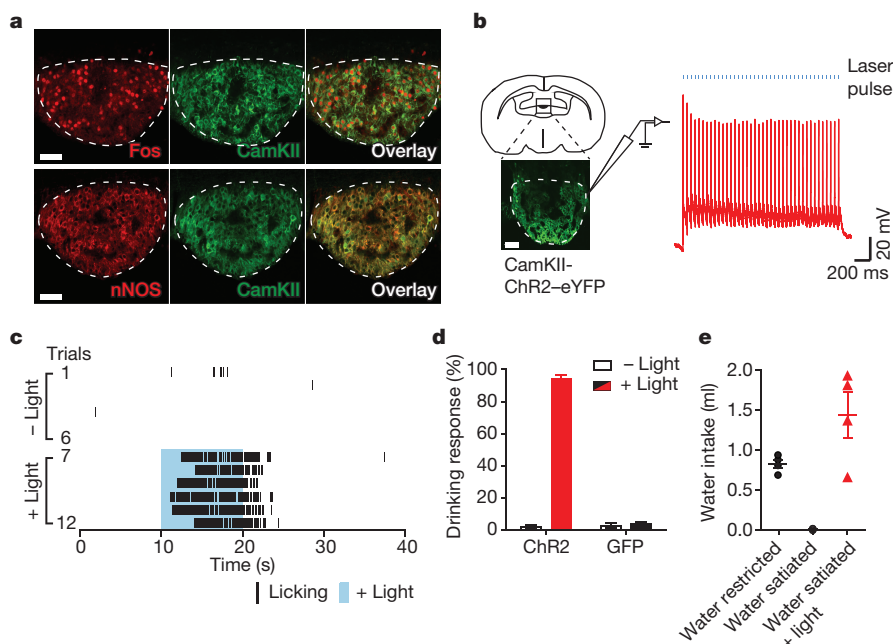
Figure 1 | Activation of excitatory neurons in the SFO triggers immediate drinking behaviour. a, Water-deprivation activates CamKII/nNOS-positive neurons in the SFO. Robust Fos expression was induced in the SFO after water restriction for 48 h. Shown are double immunolabelling for Fos and CamKII. Most Fos-positive neurons co-expressed CamKII (95.9 ± 0.3%, $n = 3$); also shown is the co-expression of CamKII with nNOS. These neurons are excitatory as they are marked by a VGlut2 trasgenic reporter[24] (Extended Data Fig. 2). **b**, Whole-cell patch-clamp recording from SFO CamKII-positive neurons in acute hypothalamic slices demonstrating light-induced activation of the ChR2-expressing neurons. Shown are traces of a representative neuron subjected to 40 pulses of ChR2 excitation (20 Hz; 2 ms pulses); blue bars denote the time and duration of the light stimulus. Scale bars, 50 μm. **c**, Photostimulation of CamKII-positive neurons in the SFO (trials 7–12; blue shading) triggered intense drinking; each black bar indicates an individual licking event. In the absence of light stimulation the same water-satiated animal

exhibits very sparse events of drinking (trials 1–6). **d**, Success of inducing drinking by photostimulation of the SFO. The drinking response (%) was calculated by determining the number of trials with more than five licks over the total number of trials; animals were tested for more than ten trials each (see Methods for details). The panel shows animals infected with AAV-CamKIIa-ChR2–eYFP ($n = 10$; red bar), and control mice infected with AAV-CamKIIa–GFP (green fluorescent protein) ($n = 4$; black bar); white bars indicate the responses in the absence of photostimulation (Mann–Whitney U-test $P < 0.0003$). **e**, Quantitation of the volume of water consumed within 15 min by three groups of animals: water-restricted for 48 h, water-satiated, and water-satiated but photostimulated during the test; light (20 Hz) was delivered with a regime of 30 s on and 30 s off for the entire 15 min session ($n = 4$, Mann–Whitney U-test, $P < 0.03$ for water-satiated ± light). Values are means ± s.e.m.

[1]Department of Biochemistry and Molecular Biophysics, Columbia College of Physicians and Surgeons, Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA. [2]Department of Neuroscience, Columbia College of Physicians and Surgeons, Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA. †Present address: Division of Biology and Biological Engineering 216-76, California Institute of Technology, Pasadena, California 91125, USA.
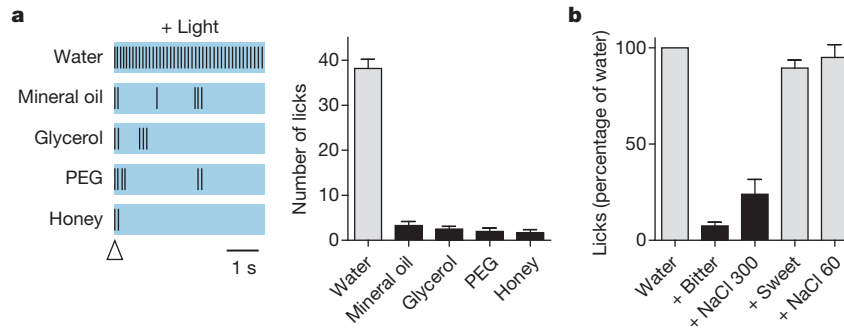
Figure 2 | CamKII-positive SFO neurons mediate thirst. Activation of CamKII-positive neurons in the SFO drives selective drinking of water. **a**, Representative raster plots illustrating licking events during a 5 s window in the presence of photostimulation; the open arrowhead indicates the first lick in each trial. The right panel shows quantification of similar data for multiple animals ($n = 6$ for honey, and 7 for others; Mann–Whitney $P < 0.002$); all animals were water-satiated. **b**, Photostimulated animals did not drink water in the presence of a bitter compound (3 μM cycloheximide; paired t-test, $P < 0.0001$), or high concentration of salt (300 mM; paired t-test, $P < 0.001$), but did so in the presence of a sweet compound (30 mM sucrose), or low salt (60 mM); data were normalized to the number of licks to water alone. Values are means ± s.e.m. ($n = 5$ animals).

intracranial injection of angiotensin, a vasoactive hormone that stimulates drinking, has been shown to activate CVO neurons in several species[9–11], and electrical stimulation of CVO nuclei increased fluid consumption in rodents[12,13].

The subfornical organ (SFO) is one of several CVO nuclei activated by thirst-inducing stimuli (for example, water-deprivation)[1,9]. This nucleus lacks the normal blood–brain barrier, and has been proposed to function as an osmolality sensor in the brain[1,14,15]. We reasoned that if we could identify a selective population of neurons in the SFO that respond to dehydration, they might provide a genetic handle to explore the neural control of thirst and water-drinking behaviour. Using Fos as a marker for neuronal activation, we found that approximately 30% of the SFO neurons were strongly labelled with Fos after a 48-h water restriction regime (no Fos expression was observed under water-satiated conditions; Extended Data Fig. 1). Notably, essentially all of the Fos-labelled cells co-expressed Ca^{2+}/calmodulin-dependent kinase II (CamKII; Fig. 1a upper panel), a known marker of excitatory neurons (see Extended Data Fig. 2), as well as neuronal nitric oxide synthase (nNOS; Fig. 1a lower panel). If these SFO neurons function as key cellular switches in the circuit that drives water consumption, then their activation should trigger water-drinking responses.

To test this hypothesis directly, we used an optogenetic approach[16,17]. We introduced ChR2 into the SFO by stereotaxic injection of an AAV-ChR2–eYFP (enhanced yellow fluorescent protein) construct under the control of the *CamkIIa*-promoter (Extended Data Fig. 3), and examined the effect of photostimulation in awake behaving animals (Figs 1b–e). Remarkably, photoactivation of the SFO CamKII-positive neurons *in vivo* triggered immediate water seeking behaviour followed by intensive drinking (Supplementary Video 1 and Fig. 1c). This response was tightly time-locked to the onset of laser stimulation, seen as long as the light stimulus was present, and could be reliably induced in over 90% of the trials (Fig. 1d). Upon termination of photostimulation the behaviour quickly ceased within a few seconds; light activation of the SFO in the absence of water had no effect on future drinking responses, even if the water was delivered just seconds after the light was switched off (Extended Data Fig. 4). Importantly, the light-induced drive to consume water was independent of the internal state of the animal as it was reliably evoked in fully water-satiated mice (Supplementary Video 2). Indeed, during a prolonged regime of laser stimulation, water-satiated mice continue to consume water avidly, and may drink nearly 8% of their body weight within 15 min; this is similar to the water consumption seen in the unstimulated animals after water restriction for 48 h (Fig. 1e). We note that light stimulation of the SFO did not induce feeding (Supplementary Video 3).

Next, we asked whether the light-induced 'thirst' is selective for water. Therefore, we assessed light-dependent fluid intake using a range of test solutions. Our results (Fig. 2a) show that the effect is highly specific for

water, with no responses to other fluids such as mineral oil, glycerol, polyethylene glycol (PEG) or even honey. Notably, light-stimulated animals refused to drink water if it contained either a bitter compound or high concentrations of salt, demonstrating that photoactivation of these SFO neurons does not bypass the natural taste-mediated functions that prevent ingestion of toxic, noxious chemicals[6] (Fig. 2b).

We identified three genetically separable, non-overlapping populations in the SFO (Fig. 3a, b and Extended Data Fig. 5): an excitatory one
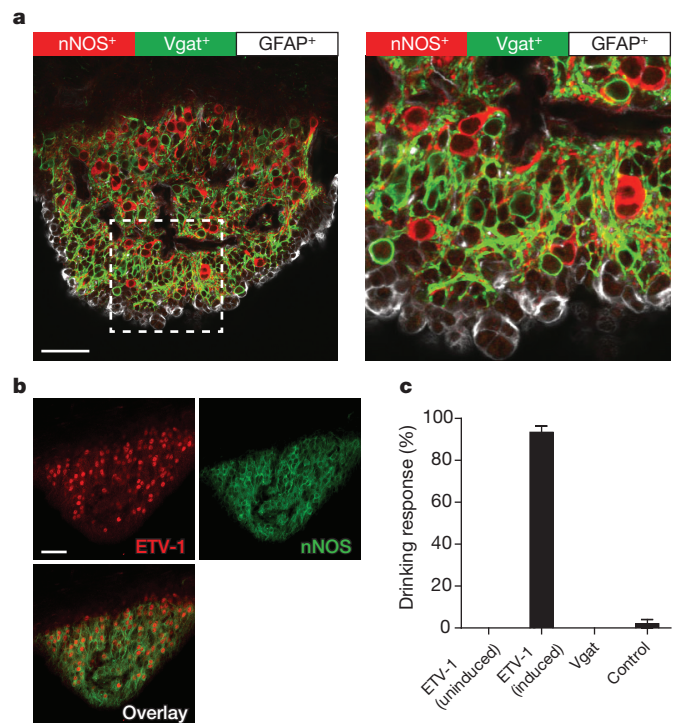
Figure 3 | Three distinct cell populations in the SFO. a, Tissue staining of the SFO from a transgenic animal expressing ChR2–eYFP in Vgat neurons (labelled with anti-GFP antibody, green) and co-labelled with anti-nNOS (red) and anti-GFAP antibodies (white); the right panel shows a magnified view illustrating the non-overlap between the three populations. **b**, ETV-1 (red) and nNOS (green) are co-expressed in most of the same neurons (>90% overlap, $n = 3$). Scale bars, 50 μm. **c**, Photostimulation of ChR2 in ETV-1-positive neurons triggers robust drinking responses in tamoxifen-induced ($n = 6$), but not uninduced animals ($n = 4$). In contrast, stimulation of ChR2 in Vgat ($n = 8$) neurons or GFAP[+] glial cells (data not shown) had no effect on drinking behaviour. Control wild-type mice infected with AAV-flex-ChR2–eYFP showed no responses to light stimulation ($n = 5$). Values are means ± s.e.m.

defined by expression of CamKII/nNOS (see Extended Data Fig. 2), and overlapping with expression of the transcription factor ETV-1, a second one defined by the expression of the vesicular GABA transporter (Vgat), and a third expressing the glial fibrillary acidic protein (GFAP; Fig. 3a). As expected, optogenetic stimulation of the ETV-1-positive neurons mimicked the effect of activating the CamKII-positive neurons and robustly triggered drinking behaviour in water-satiated animals (Fig. 3c). The *Etv1*-Cre mouse line[18] used in these experiments was tamoxifen inducible[19] (Cre-ER), and correspondingly, the behaviour was fully dependent on tamoxifen induction. Photostimulation of the other two populations did not stimulate drinking (Fig. 3c and data not shown).

Given that the CamKII/ETV-1-positive neurons provide a 'thirst-on' signal, we wondered whether one of the other cell classes might encode a 'thirst-off' signal. Indeed, activation of the Vgat-positive neurons significantly suppressed water intake in thirsty animals (>80% lick suppression); the effect was time-locked to the laser stimulation, and observed in all Vgat ChR2-expressing animals tested (Fig. 4a, b). Significantly, the suppression was as effective in thirsty animals that were actively drinking water, as it was in thirsty animals that had not yet sampled water (compare Fig. 4a, c).

If activation of Vgat-positive neurons 'quenches' thirst, then the effect should be highly specific for the motivation or drive to drink water. Thus, we examined the effect of photostimulation on salt appetite in salt-craving animals, and in sugar-intake in hungry animals (Fig. 4d, e; see also Methods). As hypothesized, activation of Vgat-positive neurons specifically extinguished the craving to consume water, but did not affect food or salt appetite. Taken together, these results substantiate Vgat-positive neurons as mediators of the 'thirst-off' signal, and demonstrate that the SFO contains selective populations of neurons mediating physiologically opposite responses to thirst.

Thirst is a fundamental physiological state representing a basic and innate response to dehydration. Earlier studies using micro-electrical stimulation and hormonal injections implicated the SFO in fluid homeostasis[9,12,13], and possibly salt appetite[20,21]. Here we have shown that the craving for water can be controlled with cell-type-specific precision in the SFO.

We used a combination of genetic and optogenetic tools in awake, behaving animals to demonstrate that the ETV-1- and Vgat-positive neurons of the SFO evoke or suppress the motivation to drink, respectively. We have shown that activation of either population instantly triggers the behaviour, be it water-seeking and drinking in normal or water-satiated animals, or strong suppression of drinking in thirsty animals; these responses are selective to water-drinking, with no effect on feeding or salt appetite. Significantly, most of the neurons in the SFO are either ETV-1-positive or Vgat-positive (Fig. 3), strongly arguing that the SFO is a dedicated brain system for thirst, functioning possibly at the interface between the physiological/internal state of the organism and the motivation to drink water. Interestingly, the ETV-1 neuronal population selectively expresses the angiotensin receptor AT1 (Extended Data Fig. 6), identifying these neurons as a possible target of angiotensin-mediated drinking responses[9–11].

In addition to the SFO, dehydration activates several other brain regions[8,15], including the organum vasculosum of the lamina terminalis[1], another hypothalamic nucleus lacking the blood–brain barrier. Notably, this nucleus has direct connections to the SFO[22,23]. Indeed, as an entry to dissect the circuit for thirst further, we surveyed the axonal projections from the ETV-1 and Vgat-expressing neurons in the SFO. Our results (Extended Data Fig. 7) show that both classes of SFO neurons project to the organum vasculosum of the lamina terminalis and the median preoptic nucleus. Interestingly, the glutamatergic neurons (that is, excitatory), unlike the GABAergic neurons, also project to the supraioptic nucleus and the paravenrtricular hypothalamic nucleus. Future physiological and behavioural studies should help reveal the role of these nodes in the neural circuitry mediating thirst, and their association with brain centres involved in other motivational states.
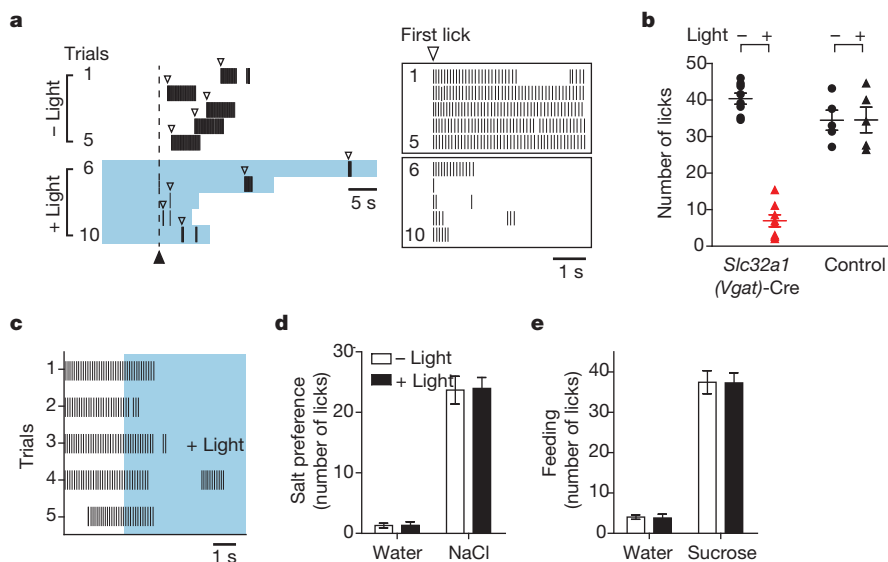
Figure 4 | Activation of Vgat-positive neurons in the SFO suppresses thirst. a, Drinking behaviour of a 24-h water-deprived animal expressing ChR2 in Vgat-positive neurons. Trials were performed in the absence (trials 1–5) or presence of photostimulation (trials 6–10). The filled arrowhead indicates the time of water presentation, and the open arrowheads mark the first lick; animals were allowed to lick for 5 s following the first lick in each trial. Light stimulation (blue shading) was started 10 s before water presentation, and maintained until the end of the 5 s licking window. The boxes on the right show an enlargement of these ten trials, each aligned to the first lick. Note the strong suppression during photostimulation. **b**, Graph quantifying the degree of suppression in animals expressing AAV-flex-ChR2–eYFP in Vgat-positive neurons of the SFO (*Slc32a1*-Cre[24]) with or without light stimulation (Mann–Whitney *U*-test, $P < 0.002$; $n = 8$). Also shown are wild-type control mice infected with the same AAV-flex-ChR2–eYFP construct ($n = 5$). Animals were tested for more than five trials each, and the total number of licks was averaged across trials. Photostimulation of the GFAP-positive population had no effect on drinking (data not shown). **c**, Activation of Vgat-positive neurons suppresses drinking behaviour even if animals were actively drinking. The plot illustrates the drinking response of a thirsty animal in five tests, before and during photostimulation (blue shading); the trials were aligned 3 s before photostimulation. **d**, **e**, Photostimulation of Vgat-positive neurons did not suppress salt appetite in salt-depleted animals (150 mM NaCl), or sugar intake in hungry animals (300 mM sucrose); values are means ± s.e.m. ($n = 7$).

1. McKinley, M. J. *et al.* The sensory circumventricular organs of the mammalian brain. *Adv. Anat. Embryol. Cell Biol.* **172,** 1–122 (2003).
2. Young, J. K. *Hunger, Thirst, Sex, and Sleep: How the Brain Controls Our Passions* (Rowman & Littlefield, 2012).
3. Sternson, S. M. Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron* **77,** 810–824 (2013).
4. Daniels, D. & Fluharty, S. J. Salt appetite: a neurohormonal viewpoint. *Physiol. Behav.* **81,** 319–337 (2004).
5. Geerling, J. C. & Loewy, A. D. Central regulation of sodium appetite. *Exp. Physiol.* **93,** 177–209 (2008).
6. Oka, Y., Butnaru, M., von Buchholtz, L., Ryba, N. J. & Zuker, C. S. High salt recruits aversive taste pathways. *Nature* **494,** 472–475 (2013).
7. Stricker, E. M. & Sved, A. F. Thirst. *Nutrition* **16,** 821–826 (2000).
8. McKinley, M. J. & Johnson, A. K. The physiological regulation of thirst and fluid intake. *News Physiol. Sci.* **19,** 1–6 (2004).
9. Fitzsimons, J. T. Angiotensin, thirst, and sodium appetite. *Physiol. Rev.* **78,** 583–686 (1998).
10. Epstein, A. N., Fitzsimons, J. T. & Simons, B. J. Drinking caused by the intracranial injection of angiotensin into the rat. *J. Physiol. (Lond.)* **200,** 98–100 (1969).
11. Sturgeon, R. D., Brophy, P. D. & Levitt, R. A. Drinking elicited by intracranial microinjection of angiotensin in the cat. *Pharmacol. Biochem. Behav.* **1,** 353–355 (1973).
12. Robertson, A., Kucharczyk, J. & Mogenson, G. J. Drinking behavior following electrical stimulation of the subfornical organ in the rat. *Brain Res.* **274,** 197–200 (1983).
13. Smith, P. M., Beninger, R. J. & Ferguson, A. V. Subfornical organ stimulation elicits drinking. *Brain Res. Bull.* **38,** 209–213 (1995).
14. Verbalis, J. G. How does the brain sense osmolality? *J. Am. Soc. Nephrol.* **18,** 3056–3059 (2007).
15. Bourque, C. W., Oliet, S. H. & Richard, D. Osmoreceptors, osmoreception, and osmoregulation. *Front. Neuroendocrinol.* **15,** 231–274 (1994).
16. Li, X. *et al.* Fast noninvasive activation and inhibition of neural and network activity by vertebrate rhodopsin and green algae channelrhodopsin. *Proc. Natl Acad. Sci. USA* **102,** 17816–17821 (2005).
17. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8,** 1263–1268 (2005).
18. Taniguchi, H. *et al.* A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71,** 995–1013 (2011).
19. Feil, R. *et al.* Ligand-activated site-specific recombination in mice. *Proc. Natl Acad. Sci. USA* **93,** 10887–10890 (1996).
20. Hiyama, T. Y., Watanabe, E., Okado, H. & Noda, M. The subfornical organ is the primary locus of sodium-level sensing by Na$_x$ sodium channels for the control of salt-intake behavior. *J. Neurosci.* **24,** 9276–9281 (2004).
21. Noda, M. & Sakuta, H. Central regulation of body-fluid homeostasis. *Trends Neurosci.* **36,** 661–673 (2013).
22. Johnson, A. K., Zardetto-Smith, A. M. & Edwards, G. L. Integrative mechanisms and the maintenance of cardiovascular and body fluid homeostasis: the central processing of sensory input derived from the circumventricular organs of the lamina terminalis. *Prog. Brain Res.* **91,** 381–393 (1992).
23. Johnson, A. K. & Gross, P. M. Sensory circumventricular organs and brain homeostatic pathways. *FASEB J.* **7,** 678–686 (1993).
24. Vong, L. *et al.* Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71,** 142–154 10.1016/j.neuron.2011. 05.028 (2011).

Author Contributions Y.O. developed the research program, designed the study, carried out the experiments, and analysed data; M.Y. performed all slice patch clamp recordings; C.S.Z. analysed data, designed experiments and together with Y.O. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.O. (yoka@caltech.edu).

METHODS

Animals. All procedures were in accordance with the US National Institutes of Health (NIH) guidelines for the care and use of laboratory animals, and were approved by the Columbia University Animal Care and Use Committee. Reported data were obtained from mice ranging from 1.5 to 4 months of age and from both genders; randomization and blinding methods were not used. C57BL/6J and transgenic animals were acquired from the Jackson Laboratory (Etv1-CreER; stock number 013048, Gfap-Cre; stock number 012886, Slc32a1 (Vgat)-Cre; stock number 016962, Ai9; stock number 007909, and Slc17a6 (Vglut2)-Cre; stock number 016963). The Etv1-CreER line was originally developed in ref. 18; Gfap-Cre line was originally developed in ref. 25; Slc32a1/Slc17a6-Cre lines were originally developed in ref. 24, Ai9 (Rosa-flex-tdTomato[26]). Animals were housed in a temperature-controlled environment with a 12-h light and 12-h dark cycle. Mice had ad libitum access to food and water except during behavioural tests. Sample sizes were chosen to allow robust statistical analysis of data; no statistical method was used to determine sample size. Representative data were chosen on the basis of at least three independent experiments.

Viral constructs. AAV viruses were prepared by the University of Pennsylvania Vector Core (AAV9.EF1α.DIO.ChR2–eYFP.WPRE, 1.07×10^{13} to 1.6×10^{13} genomic copies per millilitre; AAV9.CamKIIa.ChR2–eYFP.WPRE, 1.06×10^{13} to 1.98×10^{13} genomic copies per millilitre; AAV9.CamKIIa.GFP.WPRE, 1.71×10^{13} genomic copies per millilitre; AAV9.CB7Cl.mCherry.WPRE, 9.22×10^{12} genomic copies per millilitre; AAV9.CAG.flex.tdTomato.WPRE.bGH, 8.88×10^{12} genomic copies per millilitre).

Surgery. Adult 1.5- to 4-month-old mice were anaesthetized with ketamine and xylazine (100 mg per kg and 10 mg per kg, intraperitoneally) and placed under a stereotaxic apparatus (Narishige). During surgery, body temperature was monitored and controlled using a closed-loop heating system (FHC). Procedures for surgery and virus injection were similar to those described previously[27–29]. A small craniotomy with a diameter of less than 1 mm was performed at approximately bregma −0.55 (anterior–posterior), 0 (medial–dorsal). AAV (<40 nl total volume) was injected into the SFO by pressure injection (Nanoliter 2000, World Precision Instruments) using a pulled glass capillary at approximately 10 nl min[−1]. The coordinates for injection into the SFO were Bregma −0.55 (anterior–posterior), 0 (medial–dorsal) and 3.0 (dorsal–ventral). After injection, a 200-μm fibre bundle (ThorLabs) attached to a custom-modified ferrule (Precision Fiber Products) was placed less than 300 μm dorsal to the injection site, and permanently fixed on the skull with dental cement (Lang Dental Manufacturing). Cannulated animals were allowed to recover for at least 8 days after surgery. In Etv1-CreER mice[18], Cre-mediated ChR2 expression was induced by the injection of tamoxifen (80 mg per kg body weight for two or three times) after the recovery period. For tracing experiments, Etv1-CreER and Slc32a1-Cre mice were injected with AAV-flex-tdTomato-WPRE.bGH (<10 nl total volume). To minimize the likelihood of 'spill-over' infection in Slc32a1-Cre mice, we diluted AAV by a factor of 10 before injection.

Behavioural assays. Brief water access test. Animals were tested in a custom gustometer as described previously[6,30]. Individual trials were either 40 s (stimulation of drinking) or 60 s (suppression of drinking) duration with a minimum intertrial interval of 40 s. Trials automatically terminated 5 s after the first lick, and the number of licks in this 5-s licking window was used to quantify responses. All experiments with 24-h water restriction were performed in their home cage before testing. For experiments that extended for 48-h, animals were provided with 1 ml of water after 24 h. For salt-attraction assays (Fig. 4d), mice were injected with furosemide[6] (50 mg per kg) and were kept for 24 h with salt-deficient food (Harlan). For feeding assays (Fig. 4e), animals were food-restricted for 24 h. Data were statistically analysed using Mann–Whitney U- or two-tailed paired t-tests. After behavioural assays, animals were perfused with 4% PFA and the SFO was examined to confirm viral expression. Animals that showed no detectable viral expression in the SFO were excluded from analysis.

ChR2-mediated stimulation of drinking. Laser pulses (473 nm, 20 ms) at 20 Hz were delivered through an optic fibre bundle using a laser pulse generator (Shanghei Laser & Optics Century). The laser output was maintained at 10 mW as measured at the tip of the fibre. In each 40-s trial, animals were photostimulated for up to 20 s (10–30 s window); stimulation was terminated when the trial ended. Photostimulation was triggered manually in each trial. Animals were tested for 3–20 trials for each condition, and the number of licks was averaged across trials. Because the spout shutter automatically closed 5 s after the first lick, we excluded trials where the animals made the first lick before the experiment (photostimulation) started (0–10 s window). To analyse the efficacy of photostimulation in inducing drinking responses (Figs 1d and 3c), we determined the number of trials with more than five licks over the total number of trials. In essence, animals were photostimulated for 20 s, and we measured the number of licks during the first 5-s after they reached the spout (these were freely moving animals). If the animals exhibited more than five licks within the 5-s window, the trial was considered positive (shown as drinking

response (%)). In Fig. 1c, animals were photostimulated for 10 s with water available for the full 40-s trial. In Fig. 1e, animals were placed in a gustometer and photo-stimulation was delivered with a regime of 30 s on and 30 s off for the entire 15 min session. We measured total amount of consumed water by weighing the water bottle before and after the session.

ChR2-mediated drinking suppression. Animals were subjected to water restriction (Fig. 4a–c), salt-deprivation (Fig. 4d) or food-deprivation (Fig. 4e) for 24 h before behavioural experiments. In each 60 s trial, 473 nm laser pulses (20 ms; 10 mW at fibre tip) at 20 Hz was started 10 s before water presentation, and maintained until the end of the trial. The number of licks in a 5 s window following the first lick was analysed. Animals were tested for three to ten trials each, and the number of licks was averaged across trials.

Histology. Animals were killed with ketamine and xylazine, and perfused with 10 ml of PBS followed by 10 ml of 4% PFA in PBS (pH 7.4). Brains were dissected and post-fixed overnight in 4% PFA in PBS. Coronal brain sections (100 μm) were prepared using a vibratome (VT-1000S, Leica). After blocking with 10% FBS/0.2% Triton X-100 in PBS in the presence of 0.2% Triton X-100 for 1 h, sections were incubated with primary antibodies overnight at 4 °C. The primary antibodies (1:500 dilution) used in these studies were as follows: rabbit anti-CamKII (Abcam, ab5683), goat anti-c-Fos (Santa Cruz, SC-52G), goat anti-GFAP (Abcam, ab53554), rabbit anti-nNOS (Santa Cruz, sc-648), goat anti-nNOS (abcam ab72428), rabbit anti-ETV-1 (Abcam, ab81086) and chicken anti-GFP (Abcam, ab13970). Sections were washed twice with PBS, followed by incubation for more than 3 h with fluorophore-conjugated secondary antibodies (1:500 dilution, Jackson ImmunoResearch). Fluorescent images were acquired and processed using a confocal microscope (FV1000, Olympus). In some experiments, brain sections were counterstained with DAPI (Sigma Aldrich).

Electrophysiological recordings from the SFO neurons in acute slice preparation. Procedures for preparing acute brain slices and whole-cell recordings with optogenetic stimulations were similar to those described previously[29]. Coronal slices containing the SFO (250 μm thick) were sectioned using a vibratome (VT-1000S, Leica) in ice-cold sucrose-based solution (in mM: 213 sucrose, 26 NaHCO₃, 10 dextrose, 2.5 KCl, 2.0 MgSO₄, 2.0 CaCl₂ and 1.23 NaH₂PO₄, aerated with 95% O₂/5% CO₂). Slices were transferred to oxygenated artificial cerebrospinal fluid (composition in mM: 126 NaCl, 26 NaHCO₃, 2.5 KCl, 2 MgSO₄, 2 CaCl₂, 1.25 NaH₂PO₄ and 25 dextrose, 315 mOsm, adjusted to pH 7.4) and incubated at 32 °C for at least 40 min. SFO neurons infected with AAV-CamKII-ChR2–eYFP in vivo were visualized by differential interference contrast. Whole-cell current clamp recordings were performed at 32 °C with an Axopatch 200B amplifier and a Digidata 1440A (Molecular Devices). The patch electrode (4–6 MΩ) was filled with intracellular solution (in mM: 140 KGluconate, 3 KCl, 2 MgCl₂, 10 HEPES, 0.2 EGTA, and 2 Na₂ATP, 290 mOsm, adjusted to pH 7.2). Data were filtered at 5 kHz, sampled at 20 kHz and analysed with pClamp10 software (Molecular Devices). Photostimulation was by means of an X-Cite XLED1 (Lumen Dynamics; 470 nm, 2 ms pulses at 20 Hz).

Quantitative PCR. The SFO from Etv1-CreER/Ai9 or Slc32a1-Cre/Ai9 mice were dissected under a fluorescence microscope, ensuring minimal addition of adjoining tissue. The SFO was dissociated into single cells using Papain Dissociation System (Worthington), labelled with DAPI and the tdTomato⁺ neurons sorted using a flow cytometer (MoFlo Astrios, Beckman Coulter). RNA was extracted using a PicoPure RNA isolation kit (Applied Biosystems) and complementary DNA prepared using an Ovation RNA-seq V2 kit (Nugen). Quantitative real-time PCR used the following sets of primers: ETV-1 (5′ primer: CAAACATCCCCTTCCCACCA; 3′ primer: ATAGAACTGCCTGGGACCCT), nNOS (5′ primer: CGGGAATCAGGAGTTGCAGT; 3′ primer: CAGAGCCGTGTTCCTTTCCT), Vgat (5′ primer: TCATCGAGCTGGTGATGACG; 3′ primer: CTTGGACACGGCCTTGAGAT), AT1 (5′ primer: CAACTGCCTGAACCCTCTGT; 3′ primer: TCCACCTCAGAACAAGACGC), GAPDH (5′ primer: GGTTGTCTCCTGCGACTTCA; 3′ primer: TAGGGCCTCTCTTGCTCAGT). Data were normalized to GAPDH.

25. Garcia, A. D., Doan, N. B., Imura, T., Bush, T. G. & Sofroniew, M. V. GFAP-expressing progenitors are the principal source of constitutive neurogenesis in adult mouse forebrain. Nature Neurosci. **7,** 1233–1241 (2004).
26. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nature Neurosci. **13,** 133–140 (2010).
27. Wu, Z., Autry, A. E., Bergan, J. F., Watabe-Uchida, M. & Dulac, C. G. Galanin neurons in the medial preoptic area govern parental behaviour. Nature **509,** 325–330 (2014).
28. Aponte, Y., Atasoy, D. & Sternson, S. M. AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. Nature Neurosci. **14,** 351–355 (2011).
29. Lee, H. et al. Scalable control of mounting and attack by Esr1⁺ neurons in the ventromedial hypothalamus. Nature **509,** 627–632 (2014).
30. Chandrashekar, J. et al. The cells and peripheral representation of sodium taste in mice. Nature **464,** 297–301 (2010).
31. Ng, L, et al. An anatomic gene expression atlas of the adult mouse brain. Nature Neurosci. **12,** 356–362 (2009).
32. Yoneshima, H, et al. Er81 is expressed in a subpopulation of layer 5 neurons in rodent and primate neocortices. Neuroscience **137,** 401–412 (2006).

LETTER

The evolutionary history of lethal metastatic prostate cancer

Gunes Gundem[1], Peter Van Loo[1,2,3], Barbara Kremeyer[1], Ludmil B. Alexandrov[1], Jose M. C. Tubio[1], Elli Papaemmanuil[1], Daniel S. Brewer[4,5], Heini M. L. Kallio[6], Gunilla Högnäs[6], Matti Annala[6], Kati Kivinummi[6], Victoria Goody[1], Calli Latimer[1], Sarah O'Meara[1], Kevin J. Dawson[1], William Isaacs[7], Michael R. Emmert-Buck[8]†, Matti Nykter[6], Christopher Foster[9], Zsofia Kote-Jarai[10], Douglas Easton[11], Hayley C. Whitaker[12], ICGC Prostate UK Group‡, David E. Neal[12,13]§, Colin S. Cooper[4,10]§, Rosalind A. Eeles[10,14]§, Tapio Visakorpi[6], Peter J. Campbell[1], Ultan McDermott[1]§*, David C. Wedge[1]* & G. Steven Bova[6]§*

Cancers emerge from an ongoing Darwinian evolutionary process, often leading to multiple competing subclones within a single primary tumour[1–4]. This evolutionary process culminates in the formation of metastases, which is the cause of 90% of cancer-related deaths[5]. However, despite its clinical importance, little is known about the principles governing the dissemination of cancer cells to distant organs. Although the hypothesis that each metastasis originates from a single tumour cell is generally supported[6–8], recent studies using mouse models of cancer demonstrated the existence of polyclonal seeding from and interclonal cooperation between multiple subclones[9,10]. Here we sought definitive evidence for the existence of polyclonal seeding in human malignancy and to establish the clonal relationship among different metastases in the context of androgen-deprived metastatic prostate cancer. Using whole-genome sequencing, we characterized multiple metastases arising from prostate tumours in ten patients. Integrated analyses of subclonal architecture revealed the patterns of metastatic spread in unprecedented detail. Metastasis-to-metastasis spread was found to be common, either through *de novo* monoclonal seeding of daughter metastases or, in five cases, through the transfer of multiple tumour clones between metastatic sites. Lesions affecting tumour suppressor genes usually occur as single events, whereas mutations in genes involved in androgen receptor signalling commonly involve multiple, convergent events in different metastases. Our results elucidate in detail the complex patterns of metastatic spread and further our understanding of the development of resistance to androgen-deprivation therapy in prostate cancer.

To characterize the subclonal architecture of androgen-deprived metastatic prostate cancer, we performed whole-genome sequencing (WGS) of 51 tumours from 10 patients to an average sequencing depth of 55×, including multiple metastases from different anatomic sites in each patient and, in five cases, the prostate tumour (Supplementary Table 1). We identified a set of high-confidence substitutions, insertions/deletions, genomic rearrangements and copy number changes present in each tumour sample (Extended Data Fig. 1 and Supplementary Information, section 3). To portray the populations of tumour cells within each patient, we employed an *n*-dimensional Bayesian Dirichlet process to group clonal and subclonal mutations, that is, those mutations present in all or a fraction of tumour cells within a sample, respectively. The fraction of tumour cells carrying each mutation was calculated

from the mutant allele fraction, taking into account the tumour purity and local copy number state, as described previously[2,11]. Each of the mutations assigned to a single cluster is present in a fixed proportion of cells in each sample and hence belongs to a separate subclone, that is, a genetically distinct population of cells.

By plotting the cancer cell fractions of mutations from pairs of samples, we determined the clonal relationship between the constituent subclones and found evidence for polyclonal seeding of metastases, the most striking example of which is seen in patient A22 (Fig. 1). Each of the plots in Fig. 1a contains a cluster of mutations at (1,1), indicative of truncal mutations that were present in the most recent common ancestor of both metastases. However, in many of the plots, there are additional clusters at subclonal proportions in both samples plotted. For example, the cluster of mutations indicated by the purple circles in Fig. 1a are present in 40% of cells in A22-G, 62% of cells in A22-H, 37% of cells in A22-J and 92% of cells in A22-K. A metastasis seeded by a single cell must carry a set of mutations present in all tumour cells, representing the complement of lesions in that founding cell. In some cases, this set of mutations will be subclonal in the originating site. However, mutation clusters present subclonally in two or more samples can only occur as the result of multiple seeding events by two or more genotypically distinct cells. A graphic illustration of the clonal and subclonal clusters and their representation in all of the 10 samples from A22 is shown in Fig. 1b. Where one subclone is present in the same or a lower fraction of cells than a second subclone in all samples, the subclones are represented as nested ovals when required by the pigeonhole principle (Supplementary Information, section 4b). In contrast, clusters whose relative cancer cell fractions are reversed in different samples represent branching subclones and are shown as disjoint ovals. The full lineage relationship between the subclones can be depicted in the form of a phylogenetic tree whose branch lengths are proportional to the number of substitutions in the corresponding subclone (Fig. 1c).

In 5/10 cases (A34, A22, A31, A32, A24), we found clusters of mutations present subclonally across multiple metastases, suggesting that polyclonal seeding between different organ sites is a common occurrence in metastatic prostate cancer (Fig. 2). Mutations selected from these clusters (181–429 mutations per patient) were validated by deep sequencing (median coverage 471×) of additional aliquots of DNA from
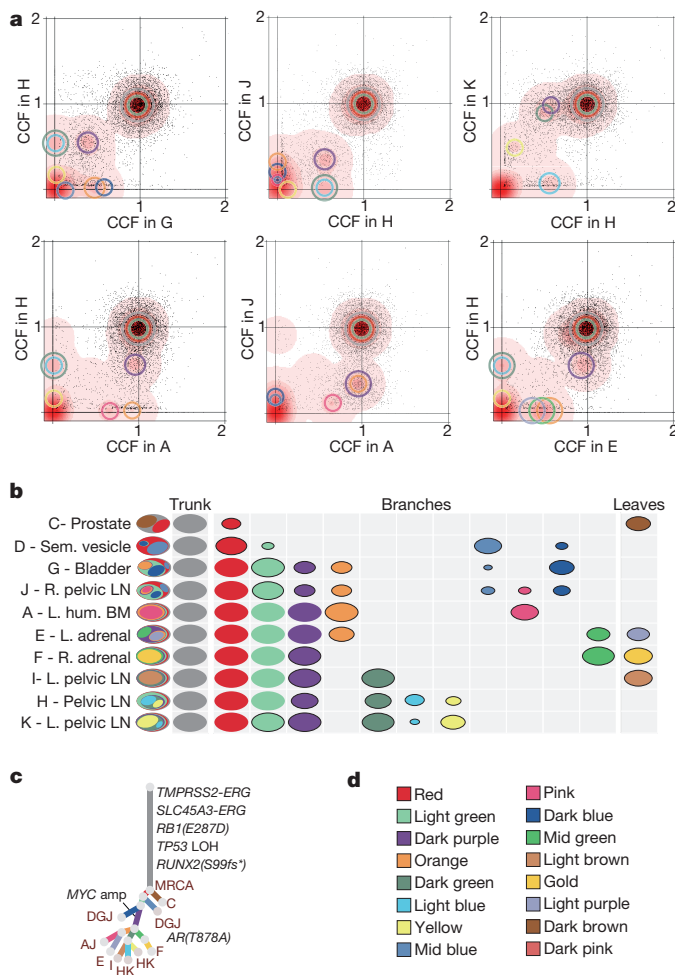
Figure 1 | n-D Dirichlet process clustering reveals widespread polyclonal seeding in A22. a, For pairs of metastases, cancer cell fractions (CCF), that is, the fraction of cancer cells within a sample containing a mutation, are plotted for all the substitutions detected in the WGS data. Red density areas off the axes and with CCF > 0 and < 1 reveal the existence of mutation clusters present at subclonal levels in more than one metastatic site. Mutation clusters for each sample are indicated with circles coloured according to the subclone they correspond to (Supplementary Table 3). The centre of each circle is positioned at the CCF values of the subclone in the two samples. The clusters at (1,1) correspond to the mutations present in all the cells in both sites (CCF = 1) while those on axes refer to sample-specific subclones. For example, light blue and dark green clusters absent from sample A are positioned on the *y* axis when H is compared to A but are moved to (0.60,0.08) and (0.60,0.88) when H is compared to K. **b,** Each subclone detected in A22 is represented as a set of colour-coded ovals across all organ sites (Supplementary Table 3). Each row represents a sample, with ovals in the far left column nested if required by the pigeonhole principle (see Supplementary Information). The area of the ovals is proportional to the CCF of the corresponding subclone. Subclonal mutation clusters are shown with solid borders. Oval plots are divided into three types: trunk (CCF = 1 in all samples), leaf (specific to a single sample) and branch (present in >1 sample and either not found in all samples or subclonal in at least one). BM, bone marrow; hum., humerus; L., left; LN, lymph node; R., right; Sem., seminal. **c,** Phylogenetic tree showing the relationships between subclones in A22. Branch lengths are proportional to the number of substitutions in each cluster. Branches are annotated with samples in which they are present and with oncogenic/putative oncogenic alterations assigned to that subclone. amp, amplification; LOH, loss of heterozygosity; MRCA, most recent common ancestor. **d,** Subclone colour key.

each WGS sample and additional metastatic and/or prostate samples, confirming these findings (Extended Data Figs 2–7, Extended Data Table 1 and Supplementary Information, section 4e).

Analysis of known driver events found in the subclones provides important insights into polyclonal spread of prostate cancer during therapy. Androgen-deprivation therapy (ADT) is the standard of care for metastatic prostate cancer and initially induces tumour regression in most patients. However, ADT inevitably results in castration-resistance through various mechanisms, including androgen receptor (*AR*) amplification, increased AR sensitivity as a result of mutation, AR phosphorylation and bypass of the AR pathway[12,13]. It is currently unknown whether castration resistance is generally acquired via a single event or more commonly appears in multiple cells independently. Two of the subclones implicated with polyclonal seeding in A22 carry different oncogenic alterations associated with ADT resistance, suggesting that clonal expansion has been driven by distinct resistance mechanisms: *MYC* amplification[14] in the purple cluster and a pathogenic *AR* substitution[15] in the mid blue cluster. Overall, in all five patients with polyclonal seeding, subclones carrying either alterations in *AR* or genes involved in *AR* signalling (such as *FOXA1*), or alternative mechanisms of castration resistance such as *MYC* amplification and *CTNNB1* mutation[16], were found to have re-seeded multiple sites. This suggests that the tumour cell populations with a significant survival advantage are not confined within the boundaries of an organ site but can successfully spread to and reseed other sites (Fig. 2).

Precise relationships between metastatic sites reveal the patterns of metastasis-to-metastasis seeding. In all seven cases for which the prostate tumour was sequenced (A10, A22, A29, A31 and A32; by targeted deep sequencing in A21 and A34), multiple metastases were more closely related to each other than any of them were to the primary tumour (Fig. 2; Extended Data Figs 2–5 and 7; Supplementary Information, section 4e). In the five cases with polyclonal seeding, this relationship resulted from multiple subclones shared subclonally by different metastases, raising the possibility of interclonal cooperativity, in agreement with recent studies using mouse models[10,17], or remodelling of metastatic niches by initial colonising prostate cancer clones, making them attractive habitats that other clones can colonise later[18]. Further, for those patients where multiple metastases from the same tissue type were analysed (A22, A34, A21), metastases located in the same tissue are more closely related than those in different tissues, as previously observed in pancreatic cancer[19]. Intriguingly, samples within close physical proximity were often more similar to each other than to more distant samples. This raises the question whether the similarity between metastases in the same tissue type arises as a result of geographical proximity or from tissue-specific seeding.

To explore further the relationships between samples, we considered the order of acquisition of mutations. Starting from the most recent common ancestor, we observe the accumulation of additional clusters of mutations representing subsequent 'selective sweeps'[20]. Phylogenetic trees give clear pictures of the order of events, allowing the creation of 'body maps' that represent emergence and movement of clones from one site to another (Fig. 3). The observed representation of subclones across different sites may be explained by two different patterns of spread: linear and branching. A22 demonstrates both patterns (Fig. 3a). The red and light green subclones are present in all metastases and indicate linear spread from the prostate to the seminal vesicle and thence to the remaining metastases. The remaining inter-site subclones have a more complex pattern demonstrating the emergence of branching lineages, each with demonstrated metastasis-to-metastasis seeding. The stepwise accumulation of clonal mutations in A21, on the other hand, displays a simple linear pattern of metastasis-to-metastasis spread (Fig. 3b). Finally, in A24, a period of sequential metastasis-to-metastasis spread was followed by parallel polyclonal spread of subclones between multiple metastases (Fig. 3c). Overall, these patterns of seeding from one metastasis to the next are seen in 8 out of the 10 patients (all but A12 and A29). We cannot formally exclude an alternative explanation for the observed patterns, that each of these metastases has seeded from an undetected subclone in the primary tumour. However, targeted re-sequencing of a subset of mutations failed to detect any such subclones,
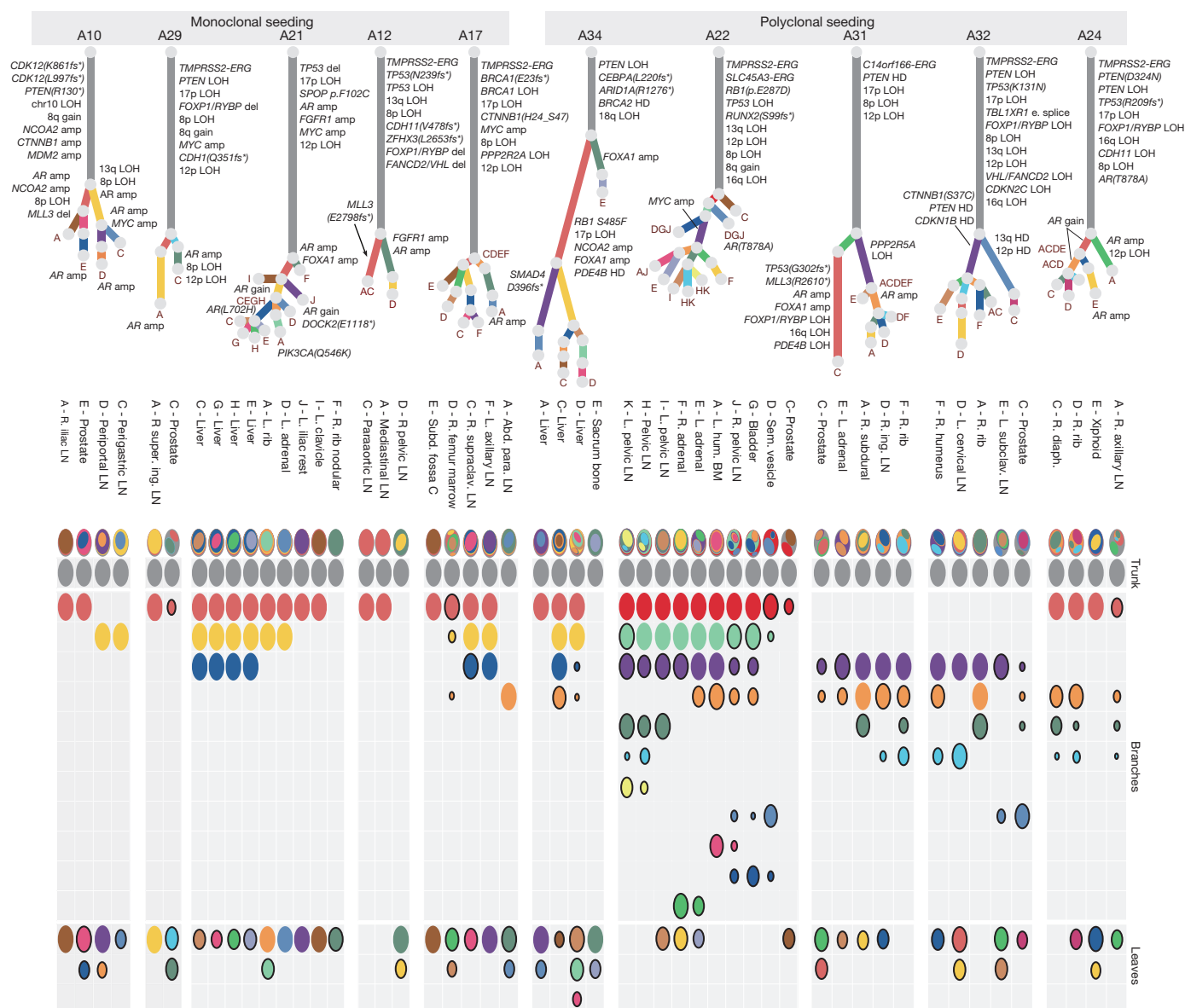
Figure 2 | Subclonal structure within 10 metastatic lethal prostate cancers. All the subclones identified in the whole-genome sequenced samples are shown as phylogenetic trees and oval plots (as described in Fig. 1). Patients with polyclonal seeding (A34, A22, A31, A32 and A24) are on the right (amp:

amplification). Abd. para., abdominal paraaortic; e. splice, essential splice; diaph., diaphragm; HD, homozygous deletion; ing., inguinal; subclav., subclavicular; super., superficial.

despite a median sequencing depth of 471× (Supplementary Information, section 4e).

Mutations found subclonally in the prostate tumour but clonally in all metastases expose the metastasizing subclone in four cases: A22, A29, A31 and A32. In each of these patients, phylogenetic reconstruction indicates that the metastases are derived from a minor subclone, encompassing fewer than 50% of the tumour cells. In three cases (A32, A10 and A34), more than one subclone from the primary tumour was involved in seeding of metastases, indicating that multiple subclones achieved metastatic potential (Supplementary Information, section 4e). In the case of A31 and A32, driver alterations that could confer selective advantage on the metastasising subclone(s) were identified (Fig. 2). In A32, both copies of *TP53* as well as one copy of *PTEN*, *RB1* and *CDKN1B*[21] were inactivated early in tumour evolution (Fig. 2). Additional aberrations occurred separately in the purple and mid blue subclones to achieve homozygous inactivation of these tumour suppressor genes via independent mechanisms (Supplementary Information, section 4e). In A31, a *PPP2R5A* deletion and an *AR* duplication occurred in the metastasising subclones (purple or orange); interestingly, the pink

cluster showed no evidence of metastatic spread, despite displaying many important oncogenic alterations including events affecting *TP53* and *MLL3* (also known as *KMT2C*; Fig. 2, Extended Data Figs 3a and 8a).

Annotation of oncogenic/putative oncogenic alterations (Supplementary Information, section 4c; Supplementary Table 2; Extended Data Table 2) on the phylogenetic trees provides some insight into the sequence of oncogenic events that take place during metastatic progression under ADT. The tumour cells in each patient share a common clonal origin (Fig. 2, grey clusters). In all patients but one (A34), this mother clone represents the largest cluster of mutations (range 40–90% of all mutations) and contains the majority of driver mutations (Figs 2 and 4a, b) similar to previous observations in pancreatic cancer[22]. In contrast, oncogenic alterations disrupting genes important for AR signalling were rarely on the trunk. All patients had at least one alteration directly affecting the *AR* locus or genes involved in *AR* signalling, with widespread heterogeneity and convergent evolution observed across multiple samples from the same patient.

In the great majority of cases, aberrations in AR signalling seem to have occurred after metastatic spread, although A21 and A24 are excep-
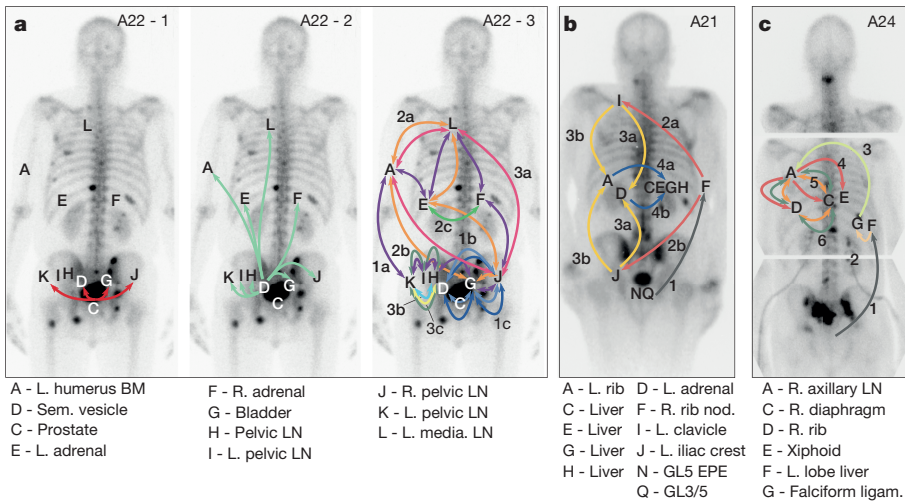
Figure 3 | Metastasis-to-metastasis seeding occurs either by a linear or by a branching pattern of spread. a–c, Body maps show the seeding of all tumour sites from A22 (**a**), A21 (**b**) and A24 (**c**). Sites shown include samples subject to targeted sequencing (A22-L, A24-F, A24-G) in addition to WGS samples. Seeding events are represented with arrows colour-coded according to Supplementary Table 3 and with double-heads when seeding could be in either direction. When the sequence of events may be ordered from the acquisition of mutations, arrows are numbered chronologically. Subclones on branching clonal lineages are labelled with the same number but with different letters, for example, 4a & 4b. See Supplementary Information section 4e for a detailed discussion of the body map in these cases. ligam., ligament. GL, Gleason grade; EPE, extrapostatic extension.

A22-1 / A22-2 / A22-3:
A - L. humerus BM
D - Sem. vesicle
C - Prostate
E - L. adrenal
F - R. adrenal
G - Bladder
H - Pelvic LN
I - L. pelvic LN
J - R. pelvic LN
K - L. pelvic LN
L - L. media. LN

A21:
A - L. rib
C - Liver
E - Liver
G - Liver
H - Liver
D - L. adrenal
F - R. rib nod.
I - L. clavicle
J - L. iliac crest
N - GL5 EPE
Q - GL3/5

A24:
A - R. axillary LN
C - R. diaphragm
D - R. rib
E - Xiphoid
F - L. lobe liver
G - Falciform ligam.

tions. The former has a large tandem duplication including the *AR* locus present in all samples, suggesting that this was an early event. The latter harbours a truncal T878A mutation, which was also detected in two additional metastases (A24-F and A24-G, interrogated by targeted se-quencing). Interestingly, a series of complex rearrangements between chromosomes 2 and X resulting in *AR* amplification was not detected in these samples (Fig. 4c). Since ADT selects for such amplification[23], it is likely that spread from the falciform ligament (A24-G) to the right axillary lymph node (A24-A) took place after ADT, which commenced 2 years and 9 months before death (Fig. 3c). Across the whole cohort, only one out of 17 *AR* amplifications was truncal, with the remainder present only in a subset of metastases. Furthermore, in five patients, AR copy number had increased on more than one occasion within the same sample (Fig. 4c and Extended Data Fig. 8), implying continuous selective pressure on the AR pathway, in line with recent reports of persistent AR signalling in castration-resistant prostate cancer[15].

Our analyses allow us to view with unprecedented clarity the geno-mic evolution of metastatic prostate cancer, from initial tumorigenesis through the acquisition of metastatic potential to the development of castration resistance. A picture emerges of a diaspora of tumour cells, sharing a common heritage, spreading from one site to another, while retaining the genetic imprint of their ancestors. After a long period of development before the most recent complete selective sweep, meta-stasis usually occurs in the form of spread between distant sites, rather than as separate waves of invasion directly from the primary tumour. This observation supports the 'seed and soil' hypothesis in which rare subclones develop metastatic potential within the primary tumour[7], rather than the theory that metastatic potential is a property of the pri-mary tumour as a whole[24,25]. Transit of cells from one host site to an-other is relatively common, either as monoclonal metastasis-to-metastasis seeding or as polyclonal seeding. Clonal diversification occurs within the constraining necessity to bypass ADT, driving distinct subclones towards a convergent path of therapeutic resistance. However, the re-sulting resistant subclones are not constrained to a single host site. Rather, a picture emerges of multiple related tumour clones compet-ing for dominance across the entirety of the host.
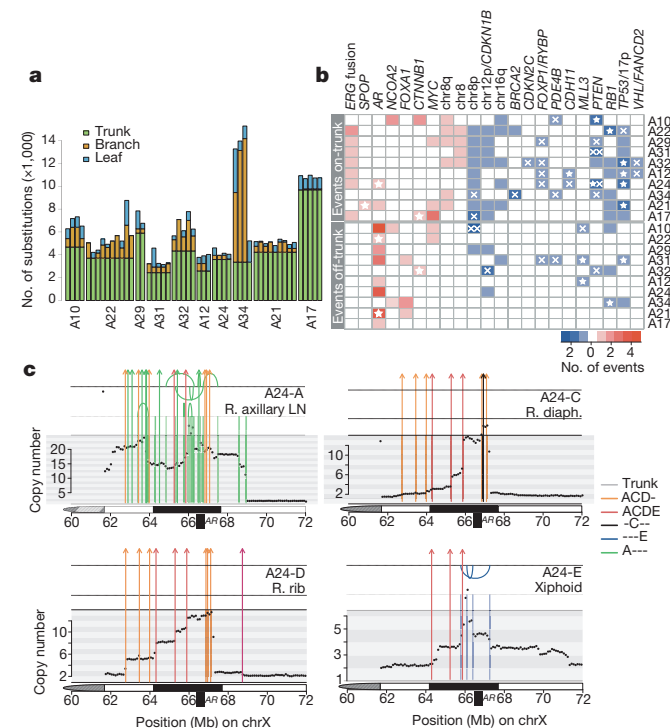
Figure 4 | Drivers of tumorigenesis are truncal while drivers of castration resistance are convergent. a, Proportion of trunk, branch and leaf mutations in each sample. **b**, Heat map of oncogenic alterations present on the trunk (top) or off the trunk (bottom), that is, on branches or leaves. Alterations in oncogenes and tumour suppressors are shown in red and blue, respectively, with shade indicating the number of events in that patient. Focal deletions and substitutions/indels are shown with crosses and stars, respectively. Double crosses indicate homozygous deletions resulting from deletions of both alleles. **c**, Continuous selective pressure on AR signalling is observed in the form of multiple rearrangements resulting in multiple copy number increases at the AR locus within the same patient. Chromosomal rearrangements are plotted on top of the genome-wide copy number for each of the 4 WGS samples from A24. Rearrangements are coloured according to the colour code in Supplementary Table 3. Arcs above and below the top vertical line indicate deletion and tandem duplication events, while arcs above and below the second vertical line are head-to-head and tail-to-tail inversions, respectively.

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
2. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
3. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
4. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
5. Gupta, G. P. & Massagué, J. Cancer metastasis: building a framework. *Cell* **127,** 679–695 (2006).
6. Poste, G. & Fidler, I. J. The pathogenesis of cancer metastasis. *Nature* **283,** 139–146 (1980).

7. Fidler, I. J. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nature Rev. Cancer* **3**, 453–458 (2003).
8. Talmadge, J. E. & Fidler, I. J. AACR centennial series: the biology of cancer metastasis: historical perspective. *Cancer Res.* **70**, 5649–5669 (2010).
9. McFadden, D. G. *et al.* Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156**, 1298–1311 (2014).
10. Cleary, A. S., Leonard, T. L., Gestl, S. A. & Gunther, E. J. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* **508**, 113–117 (2014).
11. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Commun.* **5**, 2997 (2014).
12. Karantanos, T. & Thompson, T. C. GEMMs shine a light on resistance to androgen deprivation therapy for prostate cancer. *Cancer Cell* **24**, 11–13 (2013).
13. Harris, W. P., Mostaghel, E. A., Nelson, P. S. & Montgomery, B. Androgen deprivation therapy: progress in understanding mechanisms of resistance and optimizing androgen depletion. *Nature Clin. Pract. Urol.* **6**, 76–85 (2009).
14. Bernard, D., Pourtier-Manzanedo, A., Gil, J. & Beach, D. H. Myc confers androgen-independent prostate cancer cell growth. *J. Clin. Invest.* **112**, 1724–1731 (2003).
15. Sharma, N. L. *et al.* The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer Cell* **23**, 35–47 (2013).
16. Francis, J. C., Thomsen, M. K., Taketo, M. M. & Swain, A. β-catenin is required for prostate development and cooperates with *Pten* loss to drive invasive carcinoma. *PLoS Genet.* **9**, e1003180 (2013).
17. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54–58 (2014).
18. Sun, Y. *et al.* Treatment-induced damage to the tumor microenvironment promotes prostate cancer therapy resistance through WNT16B. *Nature Med.* **18**, 1359–1368 (2012).
19. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
20. Maley, C. C. *et al.* Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res.* **64**, 3414–3427 (2004).
21. Majumder, P. K. *et al.* A prostatic intraepithelial neoplasia-dependent p27 Kip1 checkpoint induces senescence and inhibits cell proliferation and cancer progression. *Cancer Cell* **14**, 146–155 (2008).
22. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
23. Visakorpi, T. *et al. In vivo* amplification of the androgen receptor gene and progression of human prostate cancer. *Nature Genet.* **9**, 401–406 (1995).
24. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nature Genet.* **33**, 49–54 (2002).
25. Lee, Y. F. *et al.* A gene expression signature associated with metastatic outcome in human leiomyosarcomas. *Cancer Res.* **64**, 7201–7204 (2004).

METHODS

Full Methods are available in Supplementary Information.

Statistics. To determine validation rates for mutation calling, the total read depth and number of mutant reads were determined at each validation locus in validation bam files. For substitutions with a depth $\geq 20\times$, a P binomial test of statistical significance (error rate $= 1/200$) was used to calculate the probability of observing the number of mutant alleles at each locus given the total number of reads. A validation call was made where coverage of both tumour and normal samples had sufficient depth and $P < 0.05$. The validation rate for substitutions was high. On average 95% of the substitutions were absent in the matched normal and were hence called somatic (Extended Data Table 1). For indels with a depth $\geq 20\times$, the mutations were assumed to be present in the matched normal if the mutant allele burden was $\geq 1\%$ in the normal. The average validation rate for indels was 86% (Extended Data Table 1).

26. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genet.* **45,** 1134–1140 (2013).
27. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **13,** 11–22 (2010).
28. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent *SPOP, FOXA1* and *MED12* mutations in prostate cancer. *Nature Genet.* **44,** 685–689 (2012).
29. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).

LETTER

A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis

Elvin Wagenblast[1], Mar Soto[1], Sara Gutiérrez-Ángel[1], Christina A. Hartl[1], Annika L. Gable[1], Ashley R. Maceli[1], Nicolas Erard[1,2], Alissa M. Williams[1], Sun Y. Kim[1], Steffen Dickopf[1], J. Chuck Harrell[3], Andrew D. Smith[4], Charles M. Perou[3], John E. Wilkinson[5], Gregory J. Hannon[1,2] & Simon R. V. Knott[1,2]

Cancer metastasis requires that primary tumour cells evolve the capacity to intravasate into the lymphatic system or vasculature, and extravasate into and colonize secondary sites[1]. Others have demonstrated that individual cells within complex populations show heterogeneity in their capacity to form secondary lesions[2–5]. Here we develop a polyclonal mouse model of breast tumour heterogeneity, and show that distinct clones within a mixed population display specialization, for example, dominating the primary tumour, contributing to metastatic populations, or showing tropism for entering the lymphatic or vasculature systems. We correlate these stable properties to distinct gene expression profiles. Those clones that efficiently enter the vasculature express two secreted proteins, Serpine2 and Slpi, which were necessary and sufficient to program these cells for vascular mimicry. Our data indicate that these proteins not only drive the formation of extravascular networks but also ensure their perfusion by acting as anticoagulants. We propose that vascular mimicry drives the ability of some breast tumour cells to contribute to distant metastases while simultaneously satisfying a critical need of the primary tumour to be fed by the vasculature. Enforced expression of SERPINE2 and SLPI in human breast cancer cell lines also programmed them for vascular mimicry, and SERPINE2 and SLPI were overexpressed preferentially in human patients that had lung-metastatic relapse. Thus, these two secreted proteins, and the phenotype they promote, may be broadly relevant as drivers of metastatic progression in human cancer.

Until now, the most detailed studies of tumour heterogeneity have been retrospective[6–9]. For example, single cell analyses of human breast tumours have illustrated evolutionary paths of genetic diversification[9]. In such cases, genetic variation could not be associated with differences in the behaviour and capabilities of clonal populations and their specific contributions to disease. We therefore wished to complement such studies by creating an experimental model of tumour heterogeneity.

To this end, we marked individual mouse mammary carcinoma 4T1 cells with a molecular barcode via retroviral infection (Fig. 1a and Extended Data Fig. 1a). We drew from a complex mixture five different cohorts of 100,000 cells each, and introduced these orthotopically into immunocompromised recipients (NOD-SCID-$Il2rg^{-/-}$ (NSG) mice). After 24 days, primary tumours, brachial lymph nodes, blood, lungs, livers and brains were collected, and the barcode populations within each tissue were quantified (Fig. 1b and Extended Data Fig. 1f). We asked whether the subset of clones that engrafted in all samples (~1,400 clones) showed consistent behaviour in terms of contributions to aspects of disease progression across all five experiments.

Two conclusions were drawn from this analysis. First, clone abundance within the primary tumour did not correlate with abundance in circulating tumour cells (CTCs) or secondary lesions. Second, distinct

groups of clones contributed to lymph node and blood-borne metastases. Significant overlap existed between abundant clones in the blood-borne metastases and CTCs (Fig. 1b and Extended Data Fig. 1g, $P < 0.001$, hypergeometric test). However, no significant overlap was observed when comparing these sets to the prominent clones in the lymph node (Fig. 1b and Extended Data Fig. 1h). Indeed, others have reported that 20–30% of patients with distant relapse are free of axillary lymph-node metastases[10]. Thus, clonal populations within the 4T1 cell line reproducibly contribute to different aspects of disease progression.

We wished to understand the properties of these clones, which underlay their differential capabilities. We therefore established 23 clonal lines from another barcoded population (Fig. 2a and Extended Data Figs 1b and 2). All relevant barcode integration sites were mapped to ensure that no known oncogenes or tumour suppressors were altered during the barcoding process (Supplementary Information). The clones spanned a range of *in vitro* growth rates and cellular morphologies. After minimal propagation, they were pooled and orthotopically injected into NSG mice. In addition, the pool was propagated on adherent culture plates. Primary tumours and aliquots from the *in vitro* system were removed after 14 and
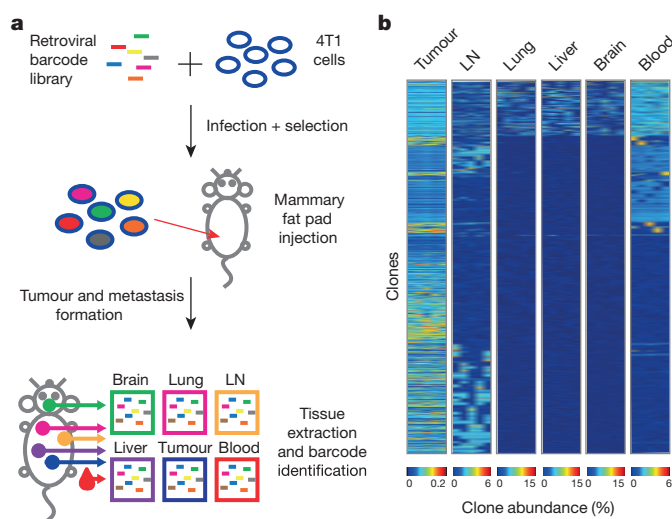
Figure 1 | Clonal analysis of 4T1 transplantation by molecular barcoding. a, Retroviral barcoding strategy for identifying clonal populations within the 4T1 cell line. **b**, Relative proportions of clones that engrafted in all animals in the lymph node (LN), lung, liver, brain and blood. Columns represent independent experiments ($n = 5$ mice). Shown are the ~1,400 clones that successfully engrafted in all animals.

24 days. In addition, at 24 days, the brachial lymph nodes, blood, lungs, livers and brains were isolated.

At 14 days, the clonal profiles of the *in vitro* samples and the primary tumours were found to be highly similar (Fig. 2b). However, at 24 days, while the *in vitro* population maintained its distribution, the primary tumour evolved along a different trajectory with clone 4T1-I dominating. Even when engrafted individually, 4T1-I showed accelerated growth between the 14- and 24-day time points, indicating that this phenotype is not dependent on clonal interactions (Extended Data Fig. 3a).

Examination of metastatic sites and CTCs showed that different clones had different capacities to contribute, and this did not correlate with their abundance in the primary tumour (Fig. 2c). Clones that were relatively less represented in the primary tumour entered the bloodstream and survived as CTCs, and a subset of these had the additional ability to colonize secondary sites. The latter clones differ still from those that colonized lymph nodes. The 4T1-T clone that dominates sites colonized by blood-borne routes was also best at forming lung metastases when injected individually (Extended Data Fig. 3b). Intravasation seemed a key gating step since intracardiac injection of the pool led to an entirely different clonal distribution in CTCs and lung metastases (Extended Data Fig. 3c).

The proclivities of each clonal line were general properties of most of their constituent cells. This was demonstrated by infection of lines 4T1-L, 4T1-E and 4T1-T with secondary, independent barcode libraries containing mCherry, allowing for populations of cells within each line to be monitored at each stage of disease (Fig. 2d and Extended Data Fig. 3d). Each secondarily barcoded line was separately pooled with the remaining lines and injected orthotopically. Similar numbers of subclones were identified for each clonal line within the tumours,

indicating that they engraft at comparable rates. A large proportion of the engrafted 4T1-E and 4T1-T subclones were able to contribute to the CTC population. Furthermore, many 4T1-T CTCs were able to extravasate and colonize the lung. The ability of the 4T1-T clone to form lung metastases was confirmed by mCherry staining of lung tissue sections from these experiments (Fig. 2e). Finally, these properties appear to be stable as they remained after the clones had been propagated for more than 20 doublings (Extended Data Fig. 3e).

To ascertain the specific drivers of this phenotype, we intersected the set of genes significantly overexpressed in clones 4T1-E and 4T1-T with a set that was found to be upregulated in lung metastases relative to matched primary tumours (Fig. 3a). Expression levels of the resultant 12 candidates were additionally examined in human patients, comparing those that did or did not relapse with lung metastases[11]. Of the 10 genes with associated patient data, the human orthologues of Serpine2 and Slpi (SERPINE2 and SLPI, respectively) emerged as the most significantly overexpressed in relapsed patients. The 4T1 cell line is used to model aggressive breast cancer subtypes such as basal, Her2 and claudin-low. Notably, it is precisely these tumour types, and not luminal cancers, that show increased SERPINE2 and SLPI expression in patients that relapse (Extended Data Fig. 4a, $P < 0.005$, Wilcoxon rank-sum test for summed expression). Additionally, a hazard analysis determined that amplified expression of these genes was significantly associated with relapse in the lung ($P < 0.0002$, Supplementary Information). Both SERPINE2 and SLPI have been implicated in breast cancer progression, however, Slpi has also been proposed to suppress tumour growth[12–16].

To validate Serpine2 and Slpi as drivers of metastatic progression, 4T1-T populations were individually infected with two short hairpin RNAs (shRNAs) targeting each gene or with a control shRNA
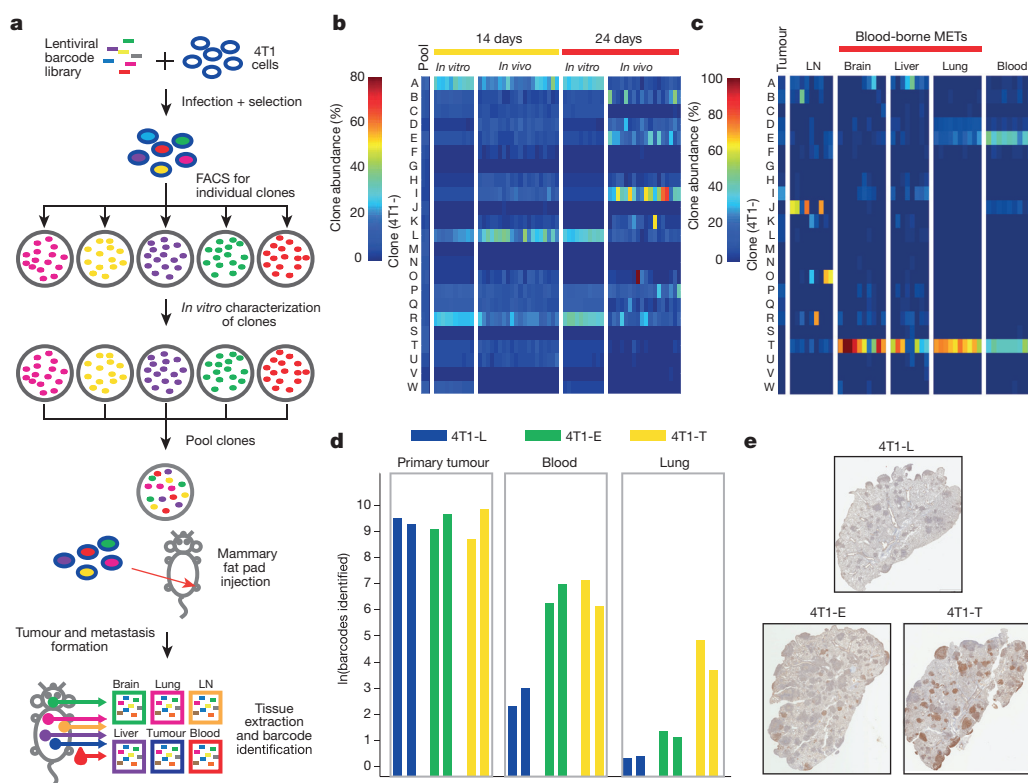
Figure 2 | Focused analysis of a subset of 4T1 clones throughout metastatic disease progression. a, Strategy for isolating individual molecularly barcoded 4T1 clones. **b,** Relative proportions of the clonal lines *in vitro* and in orthotopic primary tumours at 14 and 24 days (*in vitro* $n = 10$ cell lines, *in vivo* $n = 18$ mice). **c,** Relative proportions of clonal lines among the vascular CTCs and secondary lesions in lymph node, brain, liver and lungs in animals corresponding to the tumours extracted at 24 days in **b** ($n = 8$ mice). **d,** Subclonal

analysis of 4T1-L, 4T1-E and 4T1-T cells via secondary barcode library infection. Clonal lines were separately infected with a second barcode library, pooled with the 22 other lines and then orthotopically injected into two mice. Indicated is the number of clones identified in 100 bootstrap samples of 1 million sequence reads. **e,** Immunohistochemistry analysis for mCherry in lung metastases resulting from each of the three pooled injections discussed in **d**. mCherry is expressed in the secondary barcode library. Original magnification, ×20.

(Extended Data Fig. 1c and Supplementary Information). After orthotopic injection of a pooled collection of the resultant lines, a significant depletion of the *Serpine2* and *Slpi* shRNA expressing cells in CTCs and lung metastases was observed (Extended Data Fig. 4b, $P < 0.01$, Wilcoxon rank-sum). When each shRNA-expressing line was assessed separately in a polyclonal setting, Serpine2- and Slpi-depleted 4T1-T clones were significantly decreased in their contribution to CTCs and lung metastases (Fig. 3b, $P < 0.02$, Wilcoxon rank-sum). Also, a reduction in shRNA-expressing lung nodules was observed after silencing of either gene (Extended Data Fig. 4c, d, $P < 0.01$, Wilcoxon rank-sum). Finally, when Serpine2 and Slpi were silenced in parental 4T1 cells, singly or in combination (Extended Data Fig. 1c, d), significant reductions in lung metastases were observed. Lungs corresponding to single gene knockdowns showed a ~50% reduction in metastases and those corresponding to double knockdowns showed a ~60% reduction in secondary lesions (Extended Data Fig. 4e, f, $P < 0.01$ and $P < 0.005$, respectively, Wilcoxon rank-sum).
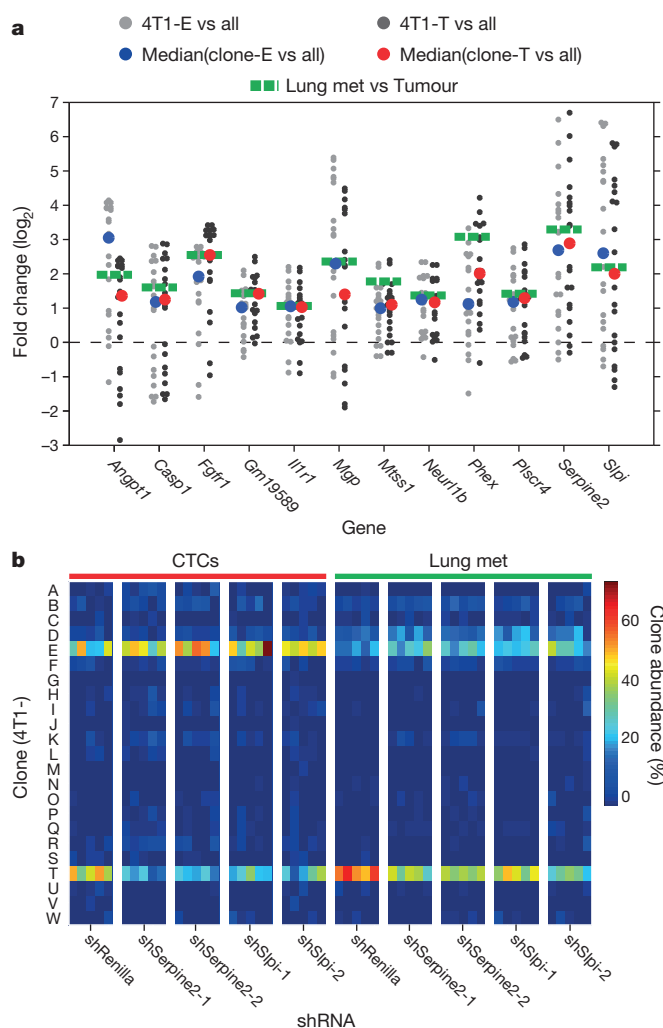
Figure 3 | Serpine2 and Slpi are regulators of intravasation into the cardiovascular system. a, Genes upregulated in the intravasating clones 4T1-E and 4T1-T. RNA sequencing (RNA-seq) was performed for all clonal lines ($n = 2$ per cell line) as well as for two pairs of matched primary tumours and lung metastases. Grey dots represent the fold change of each gene in 4T1-E (light grey) or 4T1-T (dark grey) relative to each of the other clonal lines (the median values are plotted as blue and red dots for 4T1-E and 4T1-T, respectively). The green dotted lines represent the mean fold change in the tumour and lung metastases (met) pairs. **b**, Relative proportions of clonal lines in the CTCs and lung where 4T1-T has been infected with non-targeting shRNAs and shRNAs targeting *Serpine2* and *Slpi* ($P < 0.02$, Wilcoxon rank-sum, $n = 5$ mice).

Thus, Serpine2 and Slpi probably act at the intravasation step, a hypothesis supported by the finding that intracardiac injection rescued the metastatic potential of shRNA-expressing cells (not shown). Notably, a recent study also implicated Serpins, including SERPINE2, in metastasis of breast cancer to the brain[17]. In this study, knockdown of *SERPINE2* did not block metastasis; however, cells were introduced by intracardiac injection, bypassing the requirement for increased SERPINE2 and SLPI expression according to our model.

In tumours derived from 4T1-E or 4T1-T cell lines, we observed an increase in vessels with focal loss of the endothelial cell marker CD31 (not shown). Quantitative analysis of vascular leakiness revealed that vessels within 4T1-T tumours were significantly more leaky than those of parental 4T1 and 4T1-L tumours (Extended Data Fig. 5a, b, $P < 0.05$, Wilcoxon rank-sum). Finally, silencing of Serpine2 or Slpi in these cells resulted in reduced vascular leakiness (Extended Data Fig. 5c, $P < 0.03$, Wilcoxon rank-sum).

Focal loss of CD31 in vessels and a high degree of tumour leakiness has been associated with a phenomenon termed vascular mimicry, in which tumour cells differentiate into endothelial-like cells and form extracellular-matrix-rich tubular structures to carry blood from the vasculature to hypoxic regions of the tumour[18–20]. We proposed that the propensity of 4T1-E and 4T1-T cells to intravasate was the result of a heightened capacity for vascular mimicry, placing tumour cells in direct contact with blood.

Analysis of serial sections of tumours derived from mCherry expressing 4T1-T cells revealed vast networks of structures consistent with vascular mimicry (PAS$^+$/CD31$^-$ fluid-filled channels lined by mCherry$^+$ tumour cells; Fig. 4a, b and Extended Data Fig. 6a). An equivalent number of PAS$^+$/CD31$^-$ channels were identified in 4T1-E-derived tumours. By contrast, a much lower number of such structures were identified in tumours derived from other clones (Fig. 4b). When mCherry-expressing 4T1-T cells were pooled with the remaining clones and injected orthotopically, most tumour cells surrounding the resulting PAS$^+$/CD31$^-$ vessels were mCherry-positive (Extended Data Fig. 6b).

Clones 4T1-E and 4T1-T were robust in their capacity to form tubular structures when grown on matrigel, a characteristic consistent with increased vascular mimicry in tumour cells (Extended Data Fig. 7a). In addition, Serpine2- and Slpi-depleted 4T1-T cells displayed fewer PAS$^+$/CD31$^-$ channels in their derived tumours, and formed significantly fewer tubular structures *in vitro* (Fig. 4c and Extended Data Fig. 7b, $P < 0.02$ and $P < 0.0002$, Wilcoxon rank-sum).

Enforced expression of Serpine2 and Slpi in parental 4T1 cells or non-intravasating clones 4T1-B, -F, -N and -S increased *in vitro* formation of tubular structures (with the exception of 4T1-B overexpression of Slpi, Extended Data Fig. 1e, 7c, $P < 0.03$, Wilcoxon rank-sum). For each clonal line (with the exception of 4T1-F), this was accompanied by an enhanced capacity to contribute to CTCs in the context of a heterogeneous clone mixture (for this experiment lacking 4T1-E, Extended Data Fig. 7d, $P < 0.05$, Wilcoxon rank-sum). Tumours derived from the parental 4T1 cells overexpressing Serpine2 and Slpi had significantly more PAS$^+$/CD31$^-$ channels than cells infected with an empty vector, and this was accompanied by an increase in lung metastases (Extended Data Fig. 7e–g, $P < 0.01$, Wilcoxon rank-sum, and $P < 0.05$, Friedman).

Serpine2 and Slpi also enabled vascular mimicry in human basal and/or claudin-low breast cancer lines MDA-MB-231 and -436 *in vitro* (Extended Data Fig. 8a, b, $P < 0.05$, Friedman) and *in vivo* (Fig. 4d and Extended Data Fig. 8c, $P < 0.0002$ and $P < 0.002$, respectively, Wilcoxon rank-sum). This was accompanied by an increase in lung metastatic burden (Extended Data Fig. 8d, e, $P < 0.05$, Wilcoxon rank-sum). A barcode analysis of these lines (as described for Fig. 2d) determined that this resulted from an expansion in the number of clones that were able to form CTCs (Fig. 4e and Extended Data Fig. 8f, $P < 0.04$ and $P < 0.02$, respectively, Wilcoxon rank-sum).
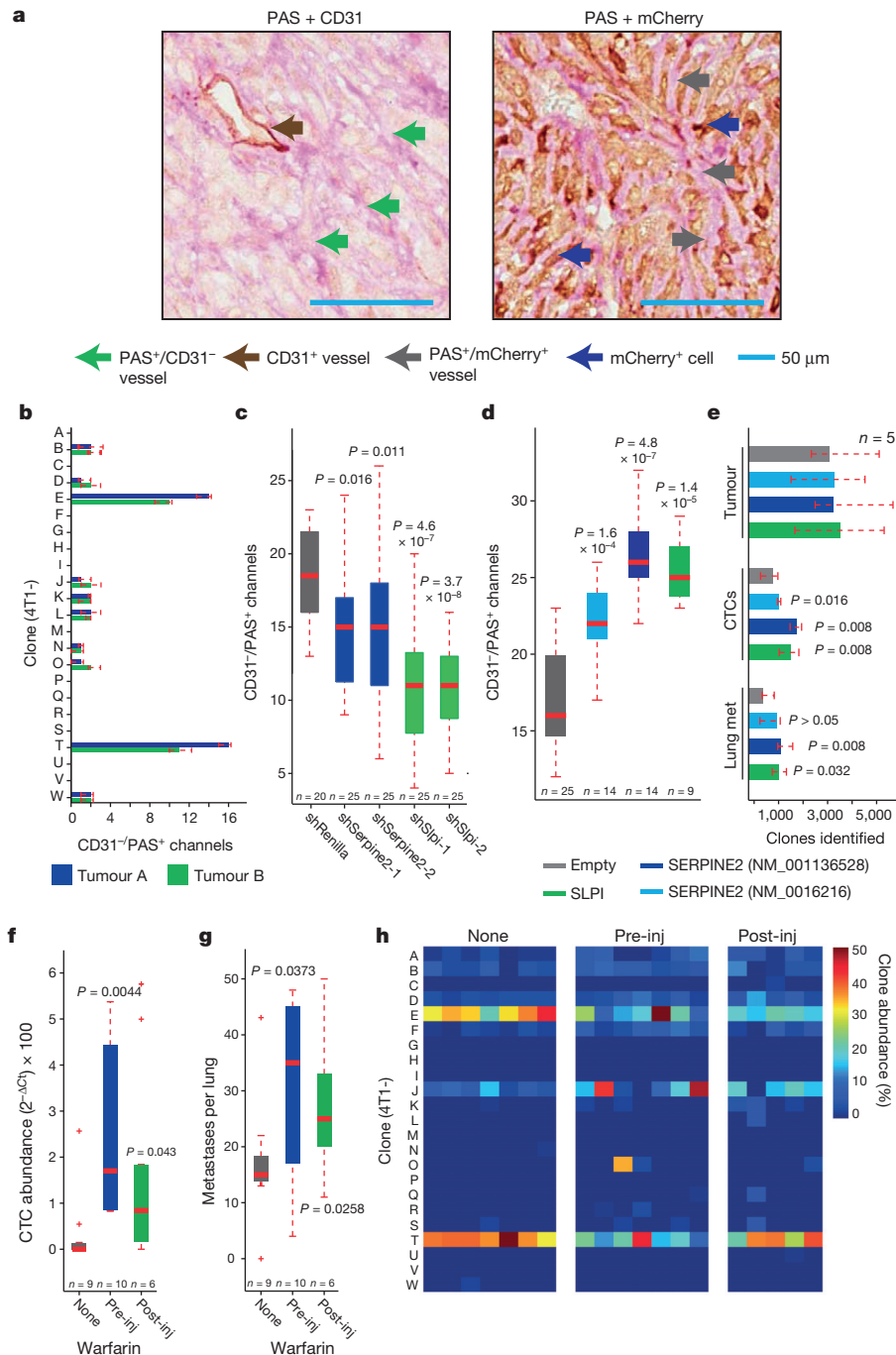
Figure 4 | Vascular mimicry drives metastatic progression. a, Serial sections of a primary tumour derived from 4T1-T cells constitutively expressing mCherry stained with PAS and either CD31 or mCherry. Brown arrow indicates CD31-positive blood vessels, green arrows CD31-negative/PAS-positive channels, blue arrows show mCherry-positive tumour cells, and grey arrows show PAS^+ channels lined by mCherry-positive tumour cells. Scale bars, 50 μm. **b,** Quantification of $CD31^-/PAS^+$ channels in tumours derived from each clonal line (median of $n = 5$ fields plotted). **c,** Quantification of $CD31^-/PAS^+$ channels in 4T1-T-derived primary tumours that have been infected with a non-targeting shRNA and shRNAs targeting *Serpine2* or *Slpi* ($P < 0.02$, Wilcoxon rank-sum, $n = 20$ fields for shRenilla and $n = 25$ fields for shSerpine2-1, shSerpine2-2, shSlpi-1 and shSlpi-2). **d,** Quantification of $CD31^-/PAS^+$ channels in MDA-MB-231-derived primary tumours cells that have been infected with an empty vector or vectors for overexpression of SERPINE2 or SLPI ($P < 0.0002$, Wilcoxon rank-sum, $n = 25$ fields for empty, $n = 14$ fields for SERPINE2 (NM_001135528 and NM_0016216) and $n = 9$ fields for SLPI). **e,** Sub-clonal analysis of MDA-MB-231 cells that have been infected with an empty vector or vectors for overexpression of SERPINE2 or

SLPI ($P < 0.04$, Wilcoxon rank-sum, $n = 5$ mice). **f,** Cardiovascular CTC abundance (measured by qPCR of the barcode vector) in animals injected orthotopically with all 23 clonal lines and administered regular drinking water or water containing $10\,\mathrm{mg\,ml}^{-1}$ warfarin ($P < 0.05$, Wilcoxon rank-sum, $n = 9$, 10 and 6 mice for mice administered no warfarin, warfarin pre-injection and warfarin post-injection, respectively). **g,** Numbers of lung metastatic nodules identified in the animals described in **f** ($P < 0.04$, Wilcoxon rank-sum, $n = 9$, 10 and 6 mice for mice administered no warfarin, warfarin pre-injection and warfarin post-injection, respectively). **h,** Relative proportions of each clone in the cardiovascular CTCs of animals that were orthotopically injected with the 23-clone pool and administered regular drinking water or water containing $10\,\mathrm{mg\,ml}^{-1}$ warfarin (either pre- or post-injection, $P < 0.002$, Wilcoxon rank-sum, $n = 7$, 7 and 5 mice for mice administered no warfarin, warfarin pre-injection and warfarin post-injection, respectively). For all box plots, the edges of the box are the twenty-fifth and seventy-fifth percentiles. The error bars extend to the values $q3 + w(q3 - q1)$ and $q1 - w(q3 - q1)$, in which w is 1.5 and $q1$ and $q3$ are the twenty-fifth and seventy-fifth percentiles, respectively.

Both Serpine2 and Slpi are anticoagulants and previous studies have reported that such factors are amplified in tumours with pronounced vascular mimicry[21]. Anticoagulants could have a role in maintaining flow in extravascular channels by preventing clotting at the vascular–extravascular interface. To test whether anticoagulants promote intravasation and metastasis, we orthotopically injected the pool of clones into warfarin-treated mice. These animals had significantly reduced levels of cleaved prothrombin factors fragments 1 and 2 (F1 and F2) in the blood, and increased vascular leakiness in their tumours (Extended Data Fig. 9a–c, $P < 0.01$ and $P < 0.000005$, Wilcoxon rank-sum). However, no significant change in PAS$^+$/CD31$^-$ vessels was observed (not shown). While tumour volumes remained stable, quantitative PCR (qPCR) for the barcode vector in whole blood revealed a significant escalation in the number of CTCs (Fig. 4f, $P < 0.05$, Wilcoxon rank-sum). This was also accompanied by an increase in lung metastatic burden (Fig. 4g, $P < 0.04$, Wilcoxon rank-sum). By contrast, when cells were injected intracardially, a significant decrease in metastases was observed, indicating that anticoagulants strongly promote intravasation (Extended Data Fig. 9d, $P < 0.05$, Wilcoxon rank-sum).

If the anticoagulant activity of Serpine2 and Slpi is important for intravasation, one would expect cells overexpressing these proteins to have less of a competitive advantage if coagulation were reduced globally within the tumour. We therefore clonally profiled the tumours, blood and lung metastases from warfarin-treated animals. Although no difference was observed between the primary tumours, clones 4T1-E and 4T1-T were reduced relatively in the CTCs and lung metastases of the warfarin-treated groups (Fig. 4h and Extended Data Fig. 9e, f, $P < 0.002$ and $P < 0.003$, Wilcoxon rank-sum). Those clones with relatively increased abundance in CTCs of warfarin-treated animals all showed some native capacity to form extravascular networks *in vivo*. These results hint that the anticoagulant action of Serpine2 and Slpi promotes extravascular network perfusion and consequently intravasation.

We have described a mouse model of breast tumour heterogeneity, which allowed us to probe the molecular basis of stable differences in the ability of clonal populations to contribute to various aspects of the disease. In this model, the ability to form CTCs, and ultimately metastases, is closely linked to the capacity for vascular mimicry. Tumour cell lined vasculature has shown a strong clinical correlation with advanced stage disease and poor clinical outcome[22]. In our model, vascular mimicry is driven by increased expression of two secreted proteins, Serpine2 and Slpi. Very little is currently known of the molecular determinants that enable vascular mimicry, and, to our knowledge, Serpine2 and Slpi are among the first validated drivers of this process. These proteins drive the formation of tubules *in vitro* and extravascular networks *in vivo*. In addition, we have shown that they probably have an additional role, by acting as anticoagulants at the vascular/extravascular interface to maintain perfusion of the tumour lined networks. Together, these properties are likely to promote the passage of red blood cells into the tumour and cancer cells into the bloodstream. Thus, our findings reveal a process that links fulfilment of the needs of the primary tumour with metastatic progression.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Vanharanta, S. & Massague, J. Origins of metastatic traits. *Cancer Cell* **24**, 410–421 (2013).
2. Miller, F. R., Miller, B. E. & Heppner, G. H. Characterization of metastatic heterogeneity among subpopulations of a single mouse mammary tumor: heterogeneity in phenotypic stability. *Invasion Metastasis* **3**, 22–31 (1983).
3. Miller, B. E., Miller, F. R., Wilburn, D. & Heppner, G. H. Dominance of a tumor subpopulation line in mixed heterogeneous mouse mammary tumors. *Cancer Res.* **48**, 5747–5753 (1988).
4. Fidler, I. J. & Kripke, M. L. Metastasis results from preexisting variant cells within a malignant tumor. *Science* **197**, 893–895 (1977).
5. Fidler, I. J. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer Res.* **38**, 2651–2660 (1978).
6. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
7. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).
8. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–1380 (2008).
9. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
10. Braun, S. *et al.* Cytokeratin-positive cells in the bone marrow and survival of patients with stage I, II, or III breast cancer. *N. Engl. J. Med.* **342**, 525–533 (2000).
11. Harrell, J. C. *et al.* Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse. *Breast Cancer Res. Treat.* **132**, 523–535 (2012).
12. Amiano, N. O. *et al.* Anti-tumor effect of SLPI on mammary but not colon tumor growth. *J. Cell. Physiol.* **228**, 469–475 (2013).
13. Fayard, B. *et al.* The serine protease inhibitor protease nexin-1 controls mammary cancer metastasis through LRP-1-mediated MMP-9 expression. *Cancer Res.* **69**, 5690–5698 (2009).
14. Martin, K. J., Patrick, D. R., Bissell, M. J. & Fournier, M. V. Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. *PLoS ONE* **3**, e2994 (2008).
15. Rosso, M. *et al.* Secretory Leukocyte Protease Inhibitor (SLPI) expression downregulates E-cadherin, induces beta-catenin re-localisation and triggers apoptosis-related events in breast cancer cells. *Biol. Cell* **106**, 308–322 (2014).
16. Sayers, K. T., Brooks, A. D., Sayers, T. J. & Chertov, O. Increased secretory leukocyte protease inhibitor (SLPI) production by highly metastatic mouse breast cancer cells. *PLoS ONE* **9**, e104223 (2014).
17. Valiente, M. *et al.* Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell* **156**, 1002–1016 (2014).
18. Maniotis, A. J. *et al.* Vascular channel formation by human melanoma cells *in vivo* and *in vitro*: vasculogenic mimicry. *Am. J. Pathol.* **155**, 739–752 (1999).
19. Hendrix, M. J., Seftor, E. A., Hess, A. R. & Seftor, R. E. Vasculogenic mimicry and tumour-cell plasticity: lessons from melanoma. *Nature Rev. Cancer* **3**, 411–421 (2003).
20. Folberg, R., Hendrix, M. J. & Maniotis, A. J. Vasculogenic mimicry and tumor angiogenesis. *Am. J. Pathol.* **156**, 361–381 (2000).
21. Ruf, W. *et al.* Differential role of tissue factor pathway inhibitors 1 and 2 in melanoma vasculogenic mimicry. *Cancer Res.* **63**, 5381–5389 (2003).
22. Cao, Z. *et al.* Tumour vasculogenic mimicry is associated with poor prognosis of human cancer patients: a systemic review and meta-analysis. *Eur. J. Cancer* **49**, 3914–3923 (2013).

Author Contributions E.W., G.J.H. and S.R.V.K. designed the experiments. E.W., M.S., S.G., C.A.H., A.L.G., A.R.M., N.E., A.M.W., S.Y.K., S.D. and S.R.V.K. performed experiments and analysed data. J.C.H. and C.M.P. assisted with human expression data. A.D.S. helped with analysis. J.E.W. performed histological analysis. E.W., G.J.H. and S.R.V.K. wrote the paper. G.J.H. and S.R.V.K. supervised the research.

METHODS

Cell culture. The mouse mammary tumour cell line 4T1 (ATCC) and any derived clonal cell lines were cultured in DMEM high glucose (Life Technologies) supplemented with 5% fetal bovine serum (FBS; Thermo Scientific), 5% fetal calf serum (FCS; Thermo Scientific), non-essential amino acids (Life Technologies) and penicillin-streptomycin (Life Technologies). Human breast tumour cell lines MDA-MB-231 and MDA-MB-436 (ATCC) were cultured in DMEM high glucose supplemented with 10% FBS, non-essential amino acids and penicillin-streptomycin. Human umbilical vein endothelial cells (HUVECs) (Lonza) were cultured in EBM-2 media with the EGM-2 Bulletkit (Lonza). HUVECs were used within three passages. Platinum-A (Cell BioLabs) and 239-FT (Life Technologies) packaging cell lines were cultured in DMEM high glucose supplemented with 10% FBS and penicillin-streptomycin.

Virus production. All retroviral vectors were packaged using platinum-A packaging cells. The lentiviral barcode library was packaged using 293-FT lentivirus packaging cells. Cells were plated on 15 cm adherent tissue culture plates (Corning) ~5 h before transfection at a confluency of ~70%. A transfection mixture was prepared with viral vector (75 µg), VSV-G (7.5 µg), 2 M calcium chloride (187.5 µl) (Sigma-Aldrich) and, when transfecting shRNA-containing vectors, 20 nM siRNAs targeting Pasha (200 µl) (Qiagen). The mixture was brought to 1.5 ml with H_2O and then added drop-wise to the same amount of 2× HBS while being bubbled. One litre of 2× HBS was prepared with 280 mM NaCl, 50 mM HEPES, 1.5 mM Na_2HPO_4, 12 mM dextrose and 10 mM KCl (Sigma-Aldrich), then adjusted to a pH of 7.02. After the transfection mixture was added to the HBS, vigorous bubbling continued for 30–60 s. After letting the resultant mixture stand for 15 min, it was added to the packaging cells along with 100 mM chloroquin (7.5 µl) (Sigma-Aldrich). After 14 h, media was replaced. Thirty hours after media change, virus was collected and filtered through a 0.45-µm filter (EMD Millipore) and stored at 4 °C.

Establishment of clonal cell lines. Around 30 million 4T1 cells were infected with the lentiviral barcode library (Extended Data Fig. 1b) at a multiplicity of infection (MOI) of 0.3. Single cells were sorted using the FACSAria IIU cell sorter (BD Biosciences) into 96-well plates. Clonal cell lines were minimally expanded and frozen down. The barcode of each individual clonal cell line was determined by Sanger sequencing. Forward primer: 5′-CAGAATCGTTGCCTGCACATCTTGGAAAC-3′ and reverse primer: 5′-ATCCAGAGGTTGATTGTTCCAGACGCGT-3′.

Clonal cell line proliferation rates. Proliferation assays were performed by counting viable cells over 72 h. In total, 1×10^5 cells were plated in duplicates and were counted using the MACSQuant Analyzer (Miltenyi Biotec).

Chromosomal integration site. Genomic DNA from each clone was isolated using the QIAamp DNA Blood Mini Kit (Qiagen). Choromosmal integration sites were determined using the Lentiviral Integration Site Analysis Kit (Clontech).

Pooling experiments. For clonal pooling experiments, clonal cell lines were counted in duplicates using the MACSQuant Analyzer (Miltenyi Biotec). Equal numbers of cells were pooled together for injection. A pre-injection pool was collected to validate equal representation of each clone before injection. Tumour, lung, brain, liver and brachial lymph node were collected from mice for further processing. Blood was collected through cardiac perfusion with PBS and 0.5 M EDTA, pH 8.0, was added as an anticoagulant.

Animal studies. All mouse experiments were approved by the Cold Spring Harbour Animal Care and Use Committee. Female 6–7-week-old NOD-SCID-$Il2rg^{-/-}$ (NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ, NSG) were purchased from JAX. All orthotopic injections were performed using 1×10^5 mouse mammary tumour cells re-suspended in 20 µl of a 1:1 mix of PBS and growth-factor-reduced Matrigel (BD Biosciences). For human breast cancer cells MDA-MB-231 and MDA-MB-436, 2×10^6 cells were re-suspended in 50 µl of a 1:1 mix of PBS and Matrigel. Injections were done into mammary gland 4. For intracardiac injections, 1×10^5 mouse mammary tumour cells were re-suspended in 200 µl of PBS and injected into the left cardiac ventricle. For tail-vein injections, 5×10^5 mouse mammary tumour cells were re-suspended in 100 µl PBS.

Primary tumour volume was measured using the formula $V = 1/2(L \times W^2)$, in which L is length and W is width of the primary tumour. Warfarin (10 mg l^{-1}, Sigma-Aldrich) was administered with drinking water and changed every 3 days. For all animal studies where a P value was to be reported, a minimum of five animals per condition were used. This allows for standard non-parametric tests (for example, Wilcoxon rank-sum) to detect strong effects. Animals were assigned to treatment groups through random cage selection. Animals that succumbed to tumour cell injections were excluded from analysis. No statistical methods were used to predetermine sample size.

Quantification of lung metastatic burden. The lung metastatic burden of individual clones and MDA-MB-231 cells injected into the mammary gland was evaluated in five-micrometre sections stained with a standard haematoxylin and eosin protocol. Quantification was performed using ImageJ Software (NIH) converting images to 8-bit. Upper and lower thresholds for each image were adjusted to determine total lung area and adjusted again to determine the metastatic area.

Both values were used to obtain relative metastatic areas. For all other experiments, the lung metastatic burden was evaluated by counting the number of metastatic nodules in the lung. For this, five-micrometre sections were stained with a standard haematoxylin and eosin protocol.

Barcode and shRNA analysis. Genomic DNA was isolated using phenol chloroform extraction for all tissues except blood. Genomic DNA for blood was isolated using the QIAamp DNA Blood Mini Kit (Qiagen).

The barcodes of the retroviral library (Extended Data Fig. 1a) were amplified using a one-step PCR protocol. For each sample, 96 individual PCR reactions of 200 ng of genomic DNA were carried out using KOD Polymerase (EMD Millipore). Forward primer: 5′-AATGATACGGCGACCACCGAGATCTACA-CTCTTTCCCTACACGACGCTCTTCCGATCT-3′ and reverse primer: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTC-AGACGTGTGCTCTTCCGATC-3′. The reverse primer contained a barcode (NNNNNN) that enabled multiplexing with standard Illumina Truseq chemistry and software. The PCR was carried out for 30 cycles and PCR products were purified using the PCR purification kit (Qiagen). PCR products were size selected on an E-gel SizeSelect 2% agarose gel (Life Technologies), and sequenced on the Illumina HiSeq sequencer generating 22-nucleotide single-end (SE) reads.

The barcodes of the lentiviral library (Extended Data Fig. 1b) were amplified using a two-step PCR protocol. For each sample, eight individual PCR reactions of 200 ng of genomic DNA were carried out using KOD Polymerase (EMD Millipore). Forward primer 1: 5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-CAGAATCGTTGCCTGCACATCTTGGAAAC-3′ and reverse primer 1: 5′-AC-ACTCTTTCCCTACACGACGCTCTTCCGATCTATCCAGAGGTTGATTGT-TCCAGACGCGT-3′. The first PCR was carried out for 25 cycles. PCR products were purified using the PCR purification kit (Qiagen). The second PCR was performed using 500 ng of PCR product from the first PCR. Forward primer 2: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC-GCTCTTCCGATCT-3 and reverse primer 2: 5′-CAAGCAGAAGACGGCAT-ACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT-C-3′. The reverse primer contained a barcode (NNNNNN) that enabled multiplexing with standard Illumina Truseq chemistry and software. The second PCR was carried out for 25 cycles and PCR products were again purified using the PCR purification kit (Qiagen). PCR products were size selected on an E-gel SizeSelect 2% agarose gel (Life Technologies), and sequenced on the Illumina HiSeq sequencer generating 22-nucleotide single-end (SE) reads.

For Figs 1, 2d, 4e and Extended Data Fig. 8f, the vector library was sequenced at high depth. For each experiment the corresponding fastq file was aligned to the vector library with the Bowtie software, allowing three mismatches. Each experimental read was then assigned to the most abundant vector sequence that it mapped to. For Fig. 1 only sequences that were present with a count greater than or equal to five in all tumours were analysed. For Fig. 2d, the error bars represent the number of clones that were identified when 100 bootstrapped samples of 1 million reads each were processed for each of the two tumours, blood and lung samples. For Fig. 4e and Extended Data Fig. 8f, bootstrapping was also performed as described above and each sample (tumour, blood sample or pair of lungs) was assigned the median of the clones identified in the corresponding random samplings.

For Extended Data Fig. 1g, h, a mixed Gaussian model was fitted to the summed distributions described in Extended Data Fig. 1f. Abundant clones were identified as those that then subsequently clustered into the Gaussian with the larger mean.

The shRNAs were amplified using the same two-step PCR protocol as described above for the lentiviral barcode library. Forward primer 1: 5′-CAG-AATCGTTGCCTGCACATCTTGGAAAC-3′ and reverse primer 1: 5′-CTGCT-AAAGCGCATGCTCCAGACTGC-3′. Forward primer 2: 5′-AATGATACGG-CGACCACCGAGATCTACACTAGCCTGCGCACGTAGTGAAGCCACAG-ATGTA-3′ and reverse primer 2: 5′-CAAGCAGAAGACGGCATACGAGAT-NNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGCTA-AAGCGCATGCTCCAGACTGC-3′. The reverse primer contained a barcode (NNNNNN) that enabled multiplexing.

NGS libraries that failed to produce more than 5,000 sequences were excluded from any further analysis as this was taken as evidence of poor quality.

Barcode quantification. Barcode libraries were de-convoluted using the Bowtie software allowing three mismatches[23]. Barcode counts were then quantified from the resultant .sam file with a simple shell script containing the unix commands, cut, sort and uniq-c.

Isolation of matched tumour and lung metastatic cells. Tumour and lung tissue were harvested from mice injected with the pool of 23 clonal cell lines. Tissue was minced and treated in DMEM high glucose containing 1× collagenase/hyaluronidase buffer (StemCell) and 10 U DNase I (Sigma) for 1 h at 37 °C. Cells were washed in HBSS (Life Technologies) twice and then re-suspended in 4T1 cell culture media containing 60 µM 6-thioguanine. Cells were passaged for 5 days until all stromal cells died.

RNA-seq library preparation. Total RNA was purified and DNase treated using the Qiagen RNeasy Mini Kit. RNA integrity (RNA Integrity score >9) and quantity was measured on an Agilent Bioanalyzer (RNA Nano kit). The NuGEN Ovation RNA-Seq V2 protocol was carried out on 100 ng of total RNA. cDNA was fragmented using the Covaris LE220 sonicator according to the manufacturer's instruction to yield a target fragment size of 200 bp. The fragmented cDNA was subsequently processed using the NuGEN Ovation Ultralow DR Multiplex System. Two technical replicates were used per sample.

Analysis of RNA-seq data. Each sample was sequenced on the Illumina HiSeq sequencer generating 76-nucleotide single-end reads. Reads were aligned to the mm10 genome using the Bowtie-2 alignment tool under default parameters[24]. Mapped reads were then assigned to genes using HTSeq-count (using the latest version of RefSeq.gtf file for gene coordinates)[25]. Resultant counts were then normalized and compared using DESeq[26]. For a gene to be considered overexpressed it had to show an at least twofold change with a false discovery rate (FDR) < 0.05.

Analysis of clinical data. A matrix where rows correspond to genes and columns to patients was quantile normalized to ensure that each patient profile had an equivalent empirical distribution. For the analysis of SERPINE2 and SLPI in relapsed and non-relapsed patients the across-patient profiles were z-score normalized and then the values of the relapsed patients compared to the non-relapse for each tumour subtype and site of relapse combination.

shRNA knockdown and cDNA overexpression. Mouse and human cell lines were transduced with amphotropically packaged retroviruses (Extended Data Fig. 1c–e). For shRNA knockdown studies, 4T1-T cells and parental 4T1 cells were selected with 500 μg ml⁻¹ hygromycin for 1 week.

For overexpression studies, all mouse clonal cells lines were selected with 1,000 μg ml⁻¹ G418 for 1 week, the parental 4T1 cell line was selected with 600 μg ml⁻¹ G418 for 1 week. MDA-MBA-231 cells were selected with 1,500 μg ml⁻¹ G418 and MDA-MD-436 cells were selected with 1,000 μg ml⁻¹ G418 for 1 week. shSerpine2-1: 5′-TGCTGTTGACAGTGAGCGACAGGTCTTCAATCAGAT-CATATAGTGAAGCCACAGATGTATATGATCTGATTGAAGACCTGGT-GCCTACTGCCTCGGA-3′. shSerpine2-2: 5′-TGCTGTTGACAGTGAGCGA-CCAGTTCAACTCTCTGTCACTTAGTGAAGCCACAGATGTAAGTGACA-GAGAGTTGAACTGGGTGCCTACTGCCTCGGA-3′. shSlpi1-1: 5′-TGCTG-TTGACAGTGAGCGATGCGTGAATCCTGTTCCCATATAGTGAAGCCAC-AGATGTATATGGGAACAGGATTCACGCACTGCCTACTGCCTCGGA-3′. shSlpi2-2: 5′-TGCTGTTGACAGTGAGCGATCAGGCAAGATGTATGATG-CTTAGTGAAGCCACAGATGTAAGCATCATACATCTTGCCTGAGTGCC-TACTGCCTCGGA-3′.

Barcode complexity studies. Parental 4T1 cells were infected with the retroviral barcode library and were selected with 500 μg ml⁻¹ hygromycin for 1 week. 4T1-E, 4T1-L and 4T1-T cells were infected with the retroviral barcode library. 4T1-E and 4T1-L cells were selected with 1,000 μg ml⁻¹ hygromycin for 1 week. 4T1-T cells were selected with 500 μg ml⁻¹ hygromycin for 1 week.

After infection and selection of MDA-MB-231 and MDA-MB-436 cells with overexpression constructs, cells were infected with the lentiviral barcode library and selected with 1 μg ml⁻¹ puromycin for 5 days.

qRT–PCR. Total RNA was purified and DNase treated using the RNeasy Mini Kit (Qiagen). Synthesis of cDNA was performed using SuperScript III Reverse Transcriptase (Sigma). Quantitative PCR analysis was performed on the Eppendorf Mastercycler ep realplex. All signals were quantified using the ΔC_t method and were normalized to the levels of *Gapdh*.

qRT–PCR primers. Mouse *Slpi* (exon 1–2): 5′-GACTGTGGAAGGAGGCAAA-3′, 5′-GGCATTGTGGCTTCTCAAG-3′. Mouse *Slpi* (exon 3–4): 5′-CAGT-GTGACGGCAAATACAAG-3′, 5′-GCCAATGTCAGGGATCAGG-3′. Mouse *Serpine2* (exon 3–4): 5′-TCTGCCTCTGAGTCCATCA-3′, 5′-AACCGAGAC-TTCCACAAACC-3′. Mouse *Serpine2* (exon 5–6): 5′-TCATCCCTCACATCA-CTACCA-3′, 5′-CTTTCAGTGGCTCCTTCAGAT-3′. Mouse *Gapdh* (exon 2–3): 5′-AATGGTGAAGGTCGGTGTG-3′, 5′-GTGGAGTCATACTGGAAC-ATGGA-3′. Human *SLPI* (exon 1–2): 5′-TGTGGAAGGCCTCTGGAAAG-3′, 5′-TGGCACTCAGGTTTCTTGTATC-3′. Human *SERPINE2* (exon 5-6): 5′-GCCATGGTGATGAGATACGG-3′, 5′-GCACTTCAATTTCAGAGGCAT-3′. Human *GAPDH* (exon 2–3): 5′-ACATCGCTCAGACACCATG-3′, 5′-TGTAG-TTGAGGTCAATGAAGGG-3′.

mCherry analysis. For immunohistochemistry, five-micrometre sections of paraffin-embedded lungs or primary tumours were deparaffinized in xylene, rehydrated in an alcohol series and immersed in distilled water. The sections were treated with high-temperature antigen retrieval in citrate buffer (pH 6). The slides were then blocked with 2.5% ready-to-use normal horse serum from ImmPRESS Anti-Rabbit Ig (peroxidase) Polymer Detection Kit (MP-7401, Vector Laboratories) for 1 h and then incubated with primary antibody RFP Antibody Pre-adsorbed (1:200) (600-401-379, Rockland) overnight at 4 °C. After washing, the slides were incubated with secondary antibody from the previous kit

for 30 min, rinsed and developed with chromogen ImmPACT DAB Peroxidase Substrate for staining (SK-4105, Vector Laboratories) until the desired intensity was achieved. Slides were counterstained with haematoxylin and coverslipped. Sections were then scanned on the Aperio Light Field Slide Scanner (Aperio) for further quantification. Omission of the primary antibody was used as a negative control in both cases. All quantification was performed in a blinded setting.

Vascular leakage. To visualize vascular leakage in the primary tumour, 100 μl of dextran Alexa 647, 10 kDa (1 mg ml⁻¹ in PBS) (Life Technologies) were injected into mice by tail vein injection. Three minutes later, mice were perfused with 4% paraformaldehyde (PFA). After fixation, tumours were collected and placed in 4% PFA overnight at 4 °C. After this, samples were infiltrated with 20% sucrose overnight at 4 °C. Tumours were frozen in OCT compound (Sakura Finetek) and 25-μm thick sections were cut, washed, incubated with DAPI (1 mg ml⁻¹) (Sigma-Aldrich) and mounted in ProLong Gold antifade reagent (Life Technologies). Sections were examined under the LSM 780 Confocal microscope (Zeiss).

An average of 5–7 fields were taken from each sample. Images were quantified using ImageJ software (NIH). For quantifying fluorescence, the threshold of each picture was adjusted to the lowest possible value in the DAPI channel to measure total tissue area. The dextran threshold was fixed in each picture at a determined value based on the average intensity of all samples processed. The dextran-positive area was then normalized to the total tissue area in order to calculate the leakiness index. All quantification was performed in a blinded setting.

CD31 analysis. Four-micrometre sections of paraffin-embedded primary tumours were de-paraffinized in xylene, rehydrated in an alcohol series and immersed in distilled water. The sections were then treated with high-temperature antigen retrieval in citrate buffer (pH 6), blocked with 2.5% ready-to-use normal horse serum from ImmPRESS Anti-Rabbit Ig (peroxidase) Polymer Detection Kit (MP-7401, Vector Laboratories) for 1 h and incubated with primary antibody against CD31 (1:400) (28364, Abcam) overnight at 4 °C. After washing, the slides were incubated with secondary antibody from the previous kit for 30 min, rinsed, and developed with chromogen ImmPACT DAB Peroxidase Substrate for staining (SK-4105, Vector Laboratories) until the desired intensity was achieved. Slides were then stained with Periodic Acid-Schiff (PAS) Kit (Sigma) according to manufacturer's instructions. Sections were then scanned on the Aperio Light Field Slide Scanner (Aperio) for further analysis. Omission of the primary antibody was used as a negative control.

Vascular mimicry. PAS staining, haematoxylin and eosin staining, and CD31 immunohistochemistry were used to evaluate the presence and extent of mimicry as previously described[18–20]. Five random ×40 fields per tumour were scored for the number and size of areas with morphology consistent with mimicry. The criteria used was (1) PAS positive channels that contain red cells and fluid, (2) the absence of CD31 staining in these channels, and (3) the polarization of tumour cells on an indistinct or imperceptible matrix lining vascular channels with red cells and/or fluid and no evidence of endothelization or tumour cells lining vascular spaces with no evidence of a matrix. All quantification was performed in a blinded setting.

Tube formation assay. The 96-well plates were coated with 50 μl of growth-factor-reduced Matrigel (BD Biosciences) and 5×10^3 cells were re-suspended in EBM-2 media and plated in each well. All cells were plated in four replicates. Morphological studies were performed after 8 h using the Zeiss Axio Observer inverted microscope. For Extended Data Fig. 7a, the morphological analysis was performed after overnight incubation.

Prothrombing fragment 1+2 ELISA. Blood was collected from animals that were treated with warfarin by cardiac heart puncture using 3.8% sodium citrate as an anticoagulant. Samples were centrifuged at 1,000*g* for 15 min at 4 °C. Blood plasma was isolated and stored at −80 °C for further processing. Plasma samples were analysed for prothrombin fragment 1+2 using the Mouse Prothrombin Fragment 1+2 (F1+2) ELISA kit (Kamiya Biomedical Company) according to manufacturer's instructions.

qPCR for circulating tumour cells. Genomic DNA for blood was isolated using the QIAamp DNA Blood Mini Kit (Qiagen) and quantified using Prime Time qPCR assays (IDT). All samples were processed in triplicates. Each reaction consisted of 50 μl, containing 25 μl of iTaq Universal Supermix (BioRad), 2.5 μl of barcode primers and probe (primer 1: 5′-ATCCAGAGGTTGATTGTTCCAGACGCGT-3′, primer 2: 5′-CAGAATCGTTGCCTGCACATCTTGGAAAC-3′, FAM probe: 5′-/56-FAM/AAGGCTCGA/ZEN/GACGTAGTCAGACGT/3IABkFQ/-3′), 2.5 μl of housekeeping (NM_172901.2) primers and probe (primer 1: 5′-GACTTGTAAC-GGGCAGGCAGATTGTG-3′, primer 2: 5′-GAGGTGTGGGTCACCTCGACA-TC-3′, HEX probe: 5′-/5HEX/CCGTGTCGC/ZEN/TCTGAAGGGCAATAT/3IABkFQ/-3, IDT) and 20 μl of gDNA sample (100 ng). The cycling conditions were 1 cycle of denaturation at 95 °C for 3 min, followed by 40 cycles of amplification (95 °C for 15 s, 68 °C for 1 min). qPCR analysis was performed on the Eppendorf

Mastercycler ep realplex. All signals were quantified using the ΔC_t method and were normalized to the levels of the housekeeping gene.

23. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

25. Anders, S., Pyl, P. T. & Huber, W. HTSeq — A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

26. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

LETTER

SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance

Dohoon Kim[1,2,3,4,5], Brian P. Fiske[3,4,5], Kivanc Birsoy[1,2,3,4,5], Elizaveta Freinkman[1,2,3,4,5], Kenjiro Kami[6], Richard L. Possemato[1,2,3,4,5], Yakov Chudnovsky[1,2,3,4,5], Michael E. Pacold[1,2,3,4,5,7], Walter W. Chen[1,2,3,4,5], Jason R. Cantor[1,2,3,4,5], Laura M. Shelton[8], Dan Y. Gui[3,4,5], Manjae Kwon[1,4], Shakti H. Ramkissoon[7,9,10], Keith L. Ligon[7,9,10], Seong Woo Kang[1,2,3,4,5], Matija Snuderl[11], Matthew G. Vander Heiden[3,4,5,7] & David M. Sabatini[1,2,3,4,5]

Cancer cells adapt their metabolic processes to support rapid proliferation, but less is known about how cancer cells alter metabolism to promote cell survival in a poorly vascularized tumour microenvironment[1–3]. Here we identify a key role for serine and glycine metabolism in the survival of brain cancer cells within the ischaemic zones of gliomas. In human glioblastoma multiforme, mitochondrial serine hydroxymethyltransferase (SHMT2) and glycine decarboxylase (GLDC) are highly expressed in the pseudopalisading cells that surround necrotic foci. We find that SHMT2 activity limits that of pyruvate kinase (PKM2) and reduces oxygen consumption, eliciting a metabolic state that confers a profound survival advantage to cells in poorly vascularized tumour regions. GLDC inhibition impairs cells with high SHMT2 levels as the excess glycine not metabolized by GLDC can be converted to the toxic molecules aminoacetone and methylglyoxal. Thus, SHMT2 is required for cancer cells to adapt to the tumour environment, but also renders these cells sensitive to glycine cleavage system inhibition.

Many inborn disorders of amino acid metabolism lead to severe impairment of the developing nervous system, at least in part through toxic effects on neural stem cells[4,5]. As brain cancer cells with high tumorigenic potential share characteristics with neural stem cells[6], we wondered whether they might have similar metabolic vulnerabilities. To begin to test this idea, we identified a set of amino acid catabolism genes whose loss causes developmental brain toxicity (Supplementary Table 1) and identified those with elevated expression in glioma compared to normal brain (Supplementary Table 2). This analysis yielded seven genes (Fig. 1a), and we focused on glycine decarboxylase (GLDC) because its expression was also highly enriched in neural stem cells (Fig. 1a). Previous work shows that elevated GLDC expression in non-small cell lung cancer tumour initiating cells promotes oncogenesis by upregulating pyrimidine biosynthesis[7]. GLDC codes for the central component of a four-protein complex (glycine cleavage complex) that catalyses the degradation of glycine into ammonia, carbon dioxide, and a methylene unit that enters the folate pool, and its loss causes nonketotic hyperglycinaemia, a disorder that severely affects the developing brain[5,8].

Consistent with the bioinformatic analysis, GLDC protein was highly expressed in tumorigenic[9,10] glioblastoma-derived neurosphere-forming cell lines BT145 and 0308, but not in their differentiated, non-tumorigenic counterparts (Extended Data Fig. 1a–c). RNA interference-mediated inhibition of GLDC caused loss of viability and breakdown of neurospheres, but did not affect the differentiated cells (Fig. 1b, Extended Data Fig. 1d and e). GLDC suppression was also toxic to LN229 cells, an adherent glioblastoma multiforme (GBM) cell line. Thus, loss of GLDC function has toxic consequences on a subset of GBM cell lines in culture.

We hypothesized that loss of GLDC may lead to the accumulation of toxic amounts of glycine. Indeed, in LN229 cells GLDC suppression raised the levels of intracellular glycine (Fig. 1c), as has been observed in the plasma in nonketotic hyperglycinaemia[5]. Interestingly, esterified glycine, which readily crosses cellular membranes and is processed into glycine[11], caused dose-dependent toxicity to the cells while other esterified amino acids did not (Fig. 1d), and this toxicity was reduced by overexpression of GLDC (Extended Data Fig. 1f).

To understand why excess glycine may be toxic to cells, we considered possible alternative fates for glycine not degraded by GLDC, its primary route of catabolism. Based on the KEGG database, there are at least 17 metabolic enzymes that process glycine, and thus we examined whether disruption of any of these other metabolic routes may affect cell sensitivity to GLDC suppression, using a pooled short hairpin RNA approach (Extended Data Fig. 2a–c). We found that suppression of glycine C-acetyltransferase (GCAT) protects against the toxicity of GLDC knockdown (Fig. 1f, Extended Data Fig. 2c, d). GCAT is part of a pathway that interconverts glycine and threonine in the mitochondria[12,13] (Fig. 1e) via 2-amino-3-ketobutyrate, an unstable intermediate that is spontaneously decarboxylated to form the toxic pro-oxidant metabolite aminoacetone[14], which itself is readily metabolized to methylglyoxal, a toxic, highly reactive aldehyde implicated in the pathology of diabetes and other disorders[15].

This raised the possibility that the glycine that is metabolized by GCAT, instead of GLDC, can be converted to aminoacetone and methylglyoxal. Indeed, GLDC knockdown or esterified glycine overload led to aminoacetone formation in LN229 cells grown in culture or as a xenografted tumour (Fig. 1g–i, Extended Data Fig. 2e, f). GLDC knockdown also increased methylglyoxal levels, as indicated by increases in argpyrimidine, a methylglyoxal-derived advanced glycation end product (Fig. 1j, Extended Data Fig. 2g). Importantly, these changes were suppressed by silencing of GCAT (Fig. 1k, Extended Data Fig. 2i). Thus glycine accumulation is deleterious, at least in part because it is converted via GCAT to aminoacetone and methylglyoxal when not sufficiently catabolized by GLDC. Recent work shows that, in the absence of serine, large quantities of glycine can be toxic by causing a depletion of the one-carbon pool that is rescued by formate supplementation[16]. Formate does not rescue glycine toxicity under our conditions (data not shown), showing that additional toxicities from excess glycine beyond depletion of one-carbon units contribute to the deleterious effects of GLDC inhibition in these cells.

To more rigorously test the idea that GLDC inhibition impairs cell viability by causing the accumulation of glycine, we suppressed the upstream enzyme serine hydroxymethyltransferase (SHMT2) (Extended Data Fig. 3a). While SHMT2 is a mitochondrial enzyme that converts serine to glycine and acts as a key source of glycine in proliferating cells, the cytoplasmic SHMT1 enzyme does not signifi-

[1]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]The David H. Koch Institute for Integrative Cancer Research at MIT, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [4]Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA. [5]Broad Institute of Harvard and MIT, Seven Cambridge Center, Cambridge, Massachusetts 02142, USA. [6]Human Metabolome Technologies, Inc., Tsuruoka 997-0052, Japan. [7]Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [8]Human Metabolome Technologies America, Inc., Boston, Massachusetts 02134, USA. [9]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [10]Department of Pathology, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [11]Department of Pathology, NYU Langone Medical Center and Medical School, New York, New York 10016, USA.
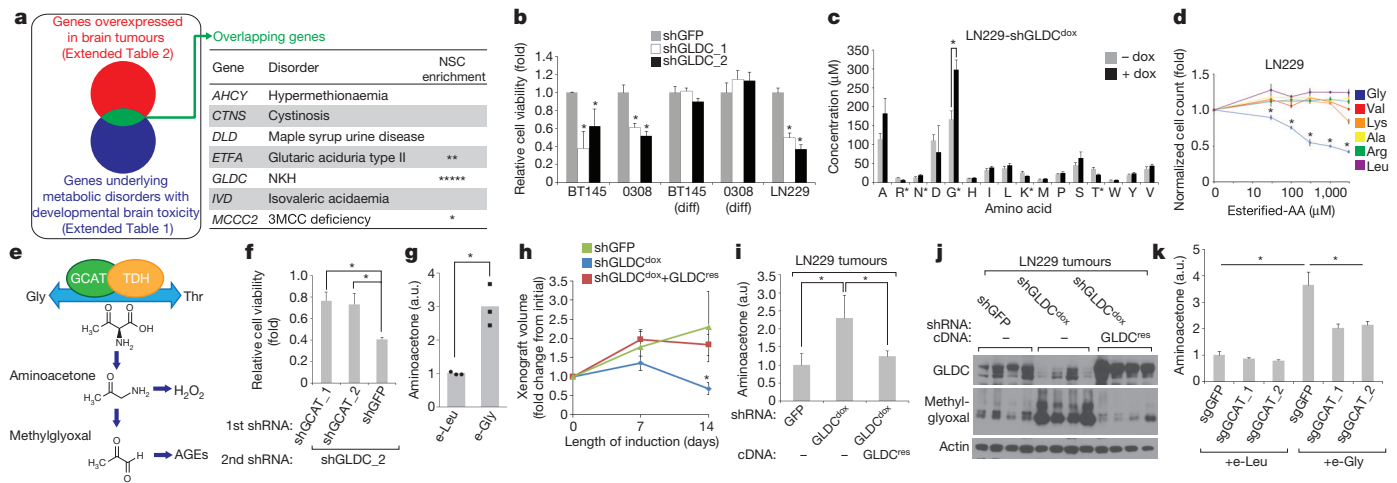
Figure 1 | GLDC is required to prevent glycine accumulation and its conversion to aminoacetone and methylglyoxal. a, Candidate gene identification scheme. Each asterisk in the 'NSC enrichment' column indicates that the given gene was significantly overexpressed (over twofold, $P < 0.05$) in neural stem cells compared to differentiated controls (Methods; total of 5 microarray studies). **b**, Viability of cells expressing the indicated shRNAs for 6 days. Values are relative to that of cells expressing shGFP. **c**, Amino acid analysis of LN229 cells with or without doxycycline (dox) induction of shGLDC[dox] for 5 days. **d**, Cell numbers following treatment with indicated doses of esterified amino acids (AA) for 5 days. Values are relative to the cell number counts of untreated controls. **e**, Diagram depicting glycine/threonine interconversion. **f**, Viability of LN229 cells first transduced with control (shGFP) or GCAT shRNAs, then transduced with shGLDC_2 shRNA for 5 days. Values are relative to that of the same cells secondarily transduced with shGFP instead of shGLDC_2. **g**, Aminoacetone levels in LN229 cells treated

with 1 mM esterified (e-) leucine or glycine for 3 days. **h**, Volumes of xenografts formed from LN229 cells expressing shGFP ($n = 5$), shGLDC[dox] ($n = 8$) or shGLDC[dox] plus shRNA-resistant GLDC ($n = 8$). Tumours were allowed to form for two weeks before doxycycline induction (Methods). Volumes are shown as relative to the starting volume (at beginning of induction) for each tumour. Error bars are s.e.m. **i**, Aminoacetone levels, normalized to tumour weight, from xenograft tumours shown in **h**, $n = 4$ per group. Error bars are s.d. **j**, Immunoblots from xenograft tumours shown in **h**. Methylglyoxal levels are indicated by argpyrimidine antibody, which recognizes proteins modified by methylglyoxal. **k**, Aminoacetone levels in cells stably transduced with Cas9 and single guide RNA against GCAT or control (GFP), and treated (4 days and 2 days before collection) with 1 mM esterified leucine or glycine. For **b**, **c**, **d**, **f**, **g** and **k**, $n = 3$ independent biological replicates; For **h** and **i**, each n described refers to the number of xenografts. For all panels. *$P < 0.05$ (Student's t-test).

cantly contribute to glycine production[17,18]. Consistent with SHMT2 functioning upstream of GLDC, suppression of SHMT2 (Extended Data Fig. 1g) decreased both net serine consumption and glycine production in LN229 cells (Fig. 2a) and completely prevented glycine

cleavage activity in isolated mitochondria as measured by [14C]CO$_2$ release (Extended Data Fig. 3a, b). Importantly, the pre-emptive knockdown of SHMT2 protected BT145, LN229 and U251 (a GBM line) cells against the detrimental effects of GLDC knockdown (Fig. 2b, c, Extended
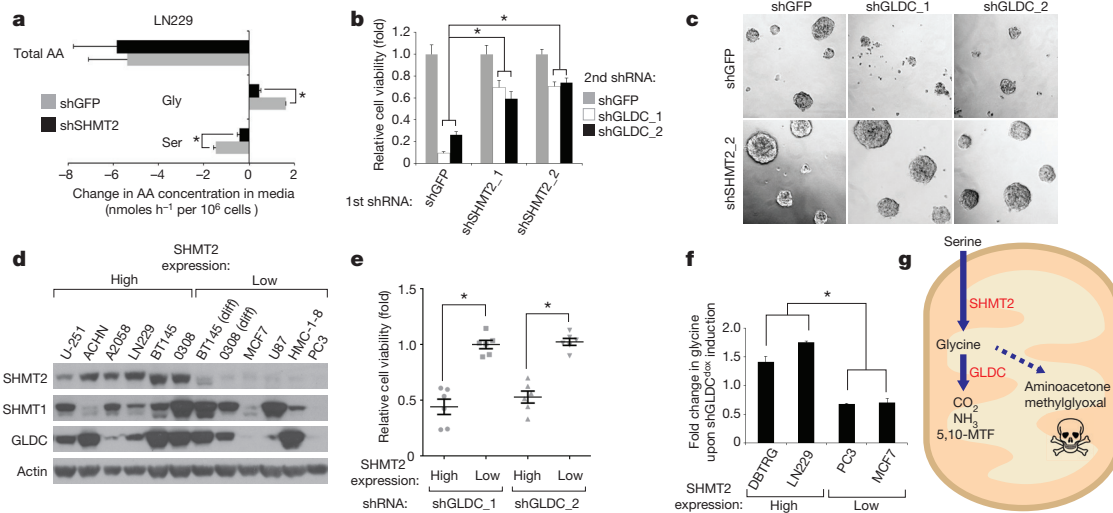
Figure 2 | SHMT2 activity renders cells liable to toxic accumulation of glycine upon GLDC loss. a, Changes in serine, glycine, and total amino acid levels over 84 h in media of LN229 cells expressing shGFP or shSHMT2_1, measured using absolute quantitative capillary electrophoresis–mass spectrometry (CE–MS) (Methods). Positive values (right of the y axis) indicate a net accumulation in the media, while negative values indicate net consumption from the media. **b**, Viability of BT145 cells first transduced with shGFP or SHMT2 shRNAs, then with shGFP or GLDC shRNAs for 5 days. Values are relative to that of cells secondarily transduced with shGFP. **c**, Representative micrographs of **b**. **d**, Immunoblots in a panel of cell lines with

high or low SHMT2 expression. **e**, Viability of cell lines in the high and low expression groups expressing shGLDC_1 and shGLDC_2 for 6–7 days. Values are relative to the viability of the same cells secondarily transduced with shGFP; individual results shown in Extended Data Fig. 3c. **f**, Glycine levels upon doxycycline-induced expression of shGLDC_2 for 5 days in different cell lines; values are relative to cells without induction; 1.0 indicates no change. **g**, Schematic of serine/glycine metabolism and cell survival in cancer cells. For **a**, **b** and **f**, $n = 3$ independent biological replicates; error bars are s.d. For **e**, each point ($n = 6$) represents a single cell line from **d**. Bars are mean ± s.e.m. For all panels, *$P < 0.05$ (Student's t-test).

Data Fig. 3d–f). These results strongly suggest that the toxicity caused by GLDC suppression is due to an accumulation of the GLDC substrate glycine instead of the depletion of 5,10-methylenetetrahydrofolate (5,10-MTF) and NADH, metabolites produced by the glycine cleavage reaction (Extended Data Fig. 3a). Furthermore, this may explain why the differentiated BT145 and 0308 cells, which express low levels of SHMT2 (Extended Data Fig. 1h, i), are insensitive to suppression of GLDC. In a panel of cancer cell lines, we found a marked correlation between SHMT2 expression levels and sensitivity to GLDC silencing (Fig. 2d and e and Extended Data Fig. 3g, h), a pattern that also matched their intracellular glycine accumulation (Fig. 2f). Collectively, these findings reveal a conditionally lethal relationship between SHMT2 and GLDC, in which SHMT2-mediated production of glycine necessitates its clearance by GLDC so as to prevent its conversion to toxic metabolites such as aminoacetone and methylglyoxal (Fig. 2g). As seen in the panel of cell lines, this relationship is probably relevant across multiple cancer cell types and is not limited to GBM cells.

In contrast to the toxic effects of GLDC knockdown, knockdown of SHMT2 did not affect the proliferation or survival of multiple cell lines under normal culture conditions (Extended Data Fig. 3i). Furthermore, SHMT2 was not necessary for the proliferation or self-renewal of neurosphere-forming cells (Extended Data Fig. 1j–l). As it seemed unlikely that cancer cells would obtain high SHMT2 expression if it did not provide a benefit, we considered that SHMT2 might have a context-dependent role and examined SHMT2 and GLDC expression in sections of human GBM tumours. In normal brains SHMT2 and GLDC expression was not detected in most cells but was at low levels in astrocytes and vessels (Fig. 3a, Extended Data Fig. 4a, d, e). In GBM tumours, however, both SHMT2 and GLDC were expressed at high levels (Fig. 3a, Extended Data Fig. 4a, d and e) that even allowed the detection of individual cancer cells migrating into the brain parenchyma (Extended Data Fig. 4b and c). Interestingly, the highest levels of SHMT2 and GLDC expression were distinct bands surrounding necrotic and acellular regions, highlighting cells of what is referred to as the pseudopalisading necrosis (Fig. 3a–c, Extended Data Fig. 4a, f, g). This feature, which is unique to glioblastomas, consists of a dense layer of "pseudopalisading" viable cells that outline an ischaemic tumour region which is thought to form upon the collapse or occlusion of an intratumoral vessel[19].
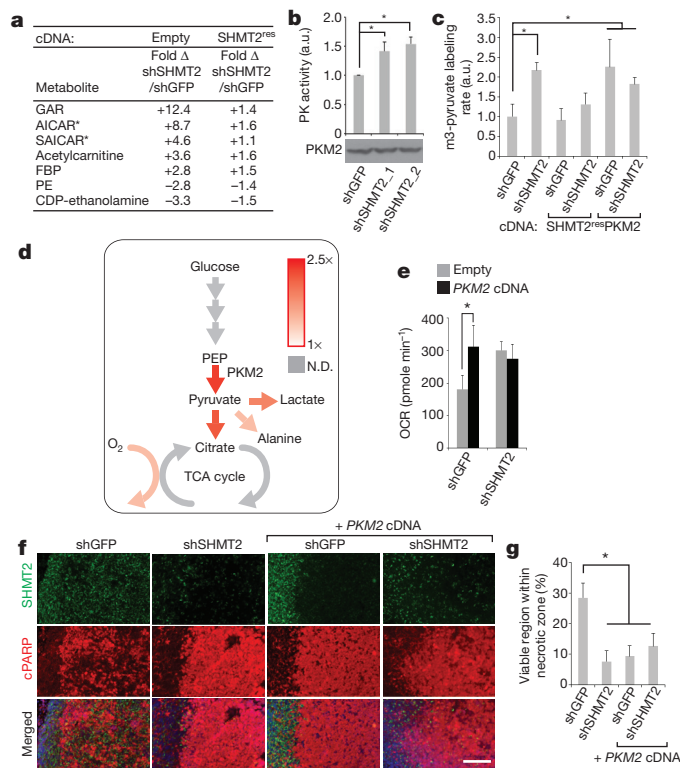
The expression of SHMT2 in ischaemic tumour zones suggested that it might have a key role in cells in environments with limited oxygen or nutrient levels. Indeed, under hypoxic conditions (0.5% oxygen), SHMT2 suppression impaired and SHMT2 overexpression enhanced LN229 cell proliferation (Fig. 3d, Extended Data Fig. 4i). As these effects were relatively modest, we set out to more closely recapitulate conditions of tumour ischaemia by using a previously described rapid xenograft model[20]. In this heterotopic model, a large bolus of cells is injected subcutaneously and the tumour collected before angiogenesis. Thus, the tumour core experiences oxygen and nutrient deprivation, which frequently results in extensive cell death, while the outermost regions of the tumour receive sufficient oxygen and nutrients and are completely viable (Fig. 3e). In such xenografts, LN229 cells, which express high levels of SHMT2 (Fig. 2d), formed a tumour with a heterogeneous central region that contained both dying cells (labelled by cleaved-PARP) and numerous 'islands' of viable cells lacking cleaved PARP (Fig. 3f, g). On the other hand, tumours formed from LN229 cells expressing a SHMT2 shRNA had a uniformly barren, cleaved-PARP immunoreactive central region that was almost completely devoid of any surviving cells. Importantly, overexpression of an RNAi-resistant *SHMT2* complementary DNA (cDNA) not only rescued the effects of SHMT2 knockdown, but also had a strong protective effect, in some cases resulting in central tumour regions that were almost entirely viable (Fig. 3f, g). While this model does not directly mimic pseudopalisading necroses, it indicates that SHMT2 expression is an important determinant of cancer cell survival within an ischaemic tumour context.
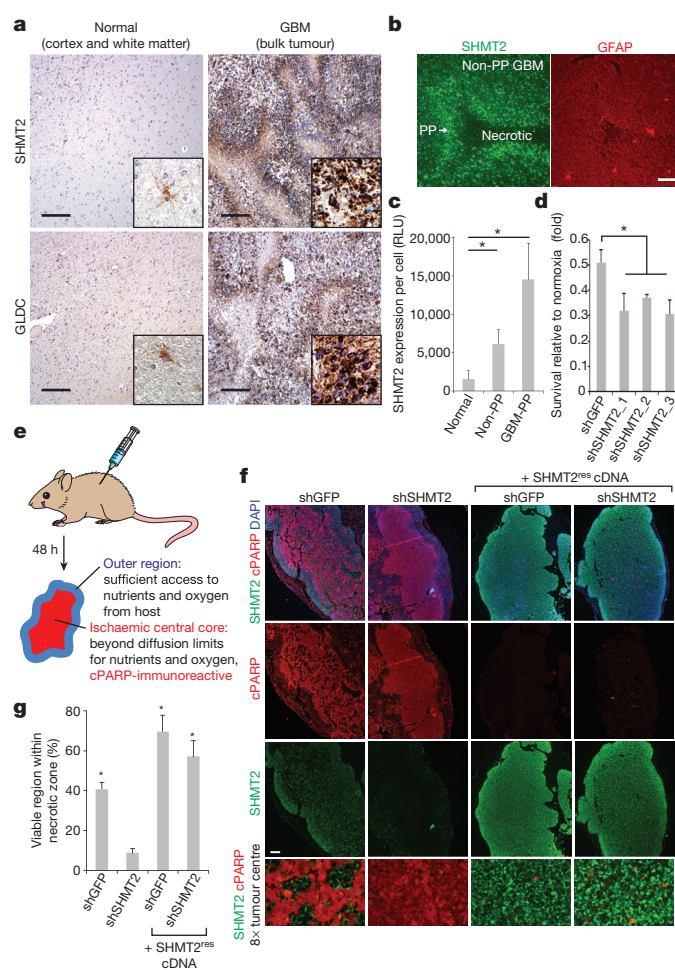
Figure 3 | SHMT2 expression provides a survival advantage in the ischaemic tumour microenvironment. a, SHMT2 and GLDC expression in normal human brains and GBM tumours. Insets are fivefold magnifications. Representative images are shown; comprehensive histological analyses are in Extended Data Fig. 4. Scale bars, 200 μm. **b**, SHMT2 immunofluorescence in the cells in the pseudopalisades (PP) and in the non-pseudopalisade GBM regions. Glial fibrillary acidic protein (GFAP), a general GBM cell marker, does not show increased signal in pseudopalisades. **c**, Quantification of SHMT2 expression, measured as fluorescence intensity per cell, in normal brain regions, non-pseudopalisade GBM regions, and pseudopalisade regions ($n = 5$ patient samples per group). Error bars are s.d. **d**, Cell number counts from LN229 cells expressing shRNAs against GFP or SHMT2 and cultured in 0.5% hypoxia for 8 days. Values are relative to the counts of the same cells cultured in parallel in normoxia; $n = 3$ independent biological replicates; error bars are s.d. **e**, Schematic of experimental design for rapid xenograft model. **f**, Representative micrographs of rapid xenograft tumours formed by LN229 cells transduced with indicated shRNAs and cDNAs, and immunostained for SHMT2 and cleaved PARP (cPARP). Bottom row shows 8× magnified, merged images of the central tumour region, displayed without DAPI channel for clarity. Scale bar, 200 μm. **g**, Quantification of the percentage of cleaved PARP-negative, viable area within the central necrotic region of xenografts of as shown in **f**. Error bars are s.e.m. (shGFP, $n = 10$; shSHMT2_1, $n = 9$; shGFP+SHMT2[res]cDNA, $n = 8$; shSHMT2_1 +SHMT2[res]cDNA, $n = 8$; each n described refers to the number of xenografts). For all panels, *$P < 0.05$ (Student's t-test).

To begin to understand why this might be, we surveyed the metabolic consequences of SHMT2 suppression in LN229 cells. Quantitative central carbon metabolism profiling revealed that, in addition to the expected accumulation of serine and depletion of glycine, SHMT2 suppression increased the levels of tricarboxylic acid (TCA) cycle intermediates and decreased those of the pentose phosphate pathway (Extended Data Fig. 5a). SHMT2 suppression also increased cellular oxygen consumption (Extended Data Fig. 5b), which may reflect increased TCA

cycle activity driving NADH into the oxidative phosphorylation pathway. Furthermore, untargeted metabolite profiling identified AICAR (5-amino-1-β-D-ribofuranosyl-imidazole-4-carboxamide), SAICAR (succinylaminoimidazolecarboxamide ribose-5-phosphate) and fructose bisphosphate (FBP) as amongst the most highly elevated metabolites (Fig. 4a and Supplementary Tables 6 and 7). The increase in the sequential intermediates SAICAR and AICAR can be explained because 10-formyltetrahydrofolate, a downstream product of SHMT2 and SHMT1 activity, is required for the conversion of AICAR to FAICAR during *de novo* purine biosynthesis (Extended Data Fig. 5d). While a

link between SHMT2 and FBP is less clear, we nonetheless noted that the suppression of SHMT2 significantly increases levels of all three known activators of pyruvate kinase isoform M2 (PKM2)—serine, FBP and SAICAR[21–23]—raising the possibility that SHMT2 antagonizes PKM2 activity by decreasing the levels of its activators.

Pyruvate kinase catalyses the conversion of phosphoenolpyruvate to pyruvate in glycolysis, and PKM2 is the isoform associated with proliferating cells[1,2]. PKM2 has regulated activity, unlike the constitutively active PKM1. Decreasing PKM2 activity can allow redistribution of glycolytic carbons in a manner advantageous for cancer cell proliferation in tumours[3,24,25], and either pharmacological PKM2 activation or PKM1 expression can impair tumour growth[25,26]. Consistent with the increase in metabolites known to activate PKM2 (Figs 2a and 4a), PKM2 activity was significantly increased in cells with suppressed SHMT2, despite no change in PKM2 protein levels (Fig. 4b, Extended Data Fig. 5c). To determine whether SHMT2 silencing induces changes in central carbon metabolism that are consistent with increased pyruvate kinase (PK) activity, we measured kinetic flux through glycolysis in live cells using ^{13}C-stable isotope labelled glucose (U-[^{13}C]glucose) (Fig. 4c, d, Extended Data Fig. 5e–h, Supplementary Tables 8 and 9). The ^{13}C labelling rate of pyruvate, the product of PKM2, was elevated in cells with suppressed SHMT2, indicating increased PKM2 activity, which was also confirmed in cells overexpressing PKM2 (Fig. 4c). By calculating the sum ^{13}C labelling of lactate, citrate and alanine, the major downstream fates of pyruvate[27], we estimate that the total pyruvate kinase flux is increased by ~70% following SHMT2 knockdown (Extended Data Fig. 5e–g). Furthermore, these changes as well as changes in metabolite levels and oxygen consumption were suppressed by overexpression of an RNAi-resistant *SHMT2* cDNA (Fig. 4c, Extended Data Fig. 5e), arguing against off-target RNAi effects. These results support a model in which SHMT2 suppression leads to increased pyruvate kinase activity and carbon flux into the TCA cycle, while cells that express high levels of SHMT2 limit PKM2 activity and flux into the TCA cycle (Fig. 4d). This may confer a survival benefit in ischaemic tumour contexts, as it has been shown that limiting pyruvate entry into TCA cycle, and thus limiting oxygen consumption, provides a survival advantage under hypoxia[28].

If the effects of SHMT2 on oxygen consumption and survival within an ischaemic microenvironment occur via suppression of PKM2 activity, then forced activation of PKM2 should antagonize these effects. Indeed, either overexpression of PKM2 or the addition of the PKM2 product pyruvate to the media increased the oxygen consumption rate in LN229 cells to the equivalent levels observed following SHMT2 knockdown (Fig. 4e and Extended Data Fig. 5i). Thus, pyruvate kinase activity may be a determinant of oxygen consumption in these cells. Furthermore, overexpression of PKM2, or the pharmacological activation of PKM2 using TEPP-46 or DASA-58 (ref. 26), reduced LN229 survival in 0.5% hypoxia to a similar extent as SHMT2 suppression (Extended Data Fig. 5j). Finally, in the rapid xenograft model, PKM2 overexpression, like SHMT2 loss, reduced the survival of LN229 cells (Fig. 4f, g). These findings support a model in which high SHMT2 expression rewires metabolism to suppress PKM2 activity and promote survival in the ischaemic tumour environment (Extended Data Fig. 5k).

In summary, we identified toxic glycine accumulation following loss of GLDC as a metabolic liability in cells expressing high levels of SHMT2. Thus, in nonketotic hyperglycinaemia, preventing endogenous glycine production via SHMT2 inhibition may be the desired route of therapy, as current treatment options targeting exogenous glycine, such as dietary restriction or plasma glycine conjugation, are largely ineffective[5].

On the other hand, SHMT2 is elevated in a subset of cancer cells and promotes changes in metabolism that allow cells to survive in an ischaemic tumour microenvironment. It is observed that hypoxia/ischaemia selects for cancer cells with increased tumorigenicity and therapy-resistance, and manifestations of tumour ischaemia, such as pseudopalisading necrosis, are associated with poor prognoses[29]. Thus, our

Figure 4 | SHMT2 elicits a PKM2-dependent metabolic rewiring that is advantageous to cancer cells in an ischaemic environment. a, Liquid chromatography–mass spectrometry (LC–MS)-based, untargeted discovery of metabolites that change in abundance following SHMT2 knockdown in LN229 cells. GAR, glycineamideribotide; FBP, fructose bisphosphate (either 1,6 or 2,6; cannot be distinguished by LC–MS); PE, phospho-ethanolamine. Differential peaks were identified and quantified as described in the Methods, and the metabolites with largest change are listed. Metabolite levels are relative and are expressed as fold change in cells transduced with shSHMT2_1 versus cells transduced with shGFP, with or without the RNAi-resistant *SHMT2* cDNA. All differences in first column are significant ($P < 0.05$). **b**, Pyruvate kinase (PK) activity assay from lysates of LN229 cells transduced with indicated shRNAs. **c**, m3-pyruvate labelling rates in LN229 cells transduced with shRNAs and cDNAs as indicated, and fed U-[^{13}C]glucose media. **d**, Summary diagram of labelling rate changes seen as a result of SHMT2 silencing. Coloured arrows indicate increased flux according to the heat map, while grey arrows indicate non-determined labelling rates. Detailed analyses are in Extended Data Fig. 5. **e**, Oxygen consumption in LN229 cells (in RPMI) expressing shGFP or shSHMT2_1 with or without *PKM2* cDNA. Error bars are s.d. ($n = 5$ technical replicates). **f**, Representative micrographs of rapid xenograft tumours formed from LN229 cells stably expressing indicated shRNAs and cDNAs, and immunostained for SHMT2 and cleaved PARP. Viable regions are oriented on the left, and the central ischaemic regions on the right. Scale bar, 100 μm. **g**, Quantification of the percentage of cleaved PARP-negative, viable area within the central necrotic region of xenografts of as shown in **f**. Error bars are s.e.m. (shGFP, $n = 10$; shSHMT2_1, $n = 8$; shGFP + *PKM2* cDNA, $n = 10$; shSHMT2_1 + *PKM2* cDNA, $n = 6$; each n described refers to the number of xenografts). For **a**, **b** and **c**, $n = 3$ independent biological replicates; error bars are s.d. For all panels, *$P < 0.05$ (Student's *t*-test).

findings raise the possibility that GLDC inhibition may be exploited to specifically target malignant and refractory subpopulations of cells expressing high levels of SHMT2.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Cantor, J. R. & Sabatini, D. M. Cancer cell metabolism: one hallmark, many faces. *Cancer Discov* **2,** 881–898 (2012).
2. Tennant, D. A., Duran, R. V. & Gottlieb, E. Targeting metabolic transformation for cancer therapy. *Nature Rev. Cancer* **10,** 267–277 (2010).
3. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324,** 1029–1033 (2009).
4. Mattson, M. P. & Shea, T. B. Folate and homocysteine metabolism in neural plasticity and neurodegenerative disorders. *Trends Neurosci.* **26,** 137–146 (2003).
5. Saudubray, J. M., Van den Berghe, G. & Walter, J. *Inborn metabolic diseases: diagnosis and treatment* 5th edn (Springer, 2012).
6. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
7. Zhang, W. C. et al. Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell* **148,** 259–272 (2012).
8. Tibbetts, A. S. & Appling, D. R. Compartmentalization of Mammalian folate-mediated one-carbon metabolism. *Annu. Rev. Nutr.* **30,** 57–81 (2010).
9. Chen, J., McKay, R. M. & Parada, L. F. Malignant glioma: lessons from genomics, mouse models, and stem cells. *Cell* **149,** 36–47 (2012).
10. Lee, J. et al. Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9,** 391–403 (2006).
11. Shikano, N. et al. Stimulation of ^{125}I-3-iodo-α-methyl-L-tyrosine uptake in Chinese hamster ovary (CHO-K1) cells by tyrosine esters. *Nucl. Med. Biol.* **37,** 189–196 (2010).
12. Dale, R. A. Catabolism of threonine in mammals by coupling of L-threonine 3-dehydrogenase with 2-amino-3-oxobutyrate-CoA ligase. *Biochim. Biophys. Acta* **544,** 496–503 (1978).
13. Tressel, T., Thompson, R., Zieske, L. R., Menendez, M. I. & Davis, L. Interaction between L-threonine dehydrogenase and aminoacetone synthetase and mechanism of aminoacetone production. *J. Biol. Chem.* **261,** 16428–16437 (1986).
14. Sartori, A. et al. Aminoacetone, a putative endogenous source of methylglyoxal, causes oxidative stress and death to insulin-producing RINm5f cells. *Chem. Res. Toxicol.* **21,** 1841–1850 (2008).
15. Kalapos, M. P. Methylglyoxal in living organisms: chemistry, biochemistry, toxicology and biological implications. *Toxicol. Lett.* **110,** 145–175 (1999).
16. Labuschagne, C. F., van den Broek, N. J., Mackay, G. M., Vousden, K. H. & Maddocks, O. D. Serine, but not glycine, supports one-carbon metabolism and proliferation of cancer cells. *Cell Rep.* **7,** 1248–1258 (2014).
17. Jain, M. et al. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **336,** 1040–1044 (2012).
18. Narkewicz, M. R., Sauls, S. D., Tjoa, S. S., Teng, C. & Fennessey, P. V. Evidence for intracellular partitioning of serine and glycine metabolism in Chinese hamster ovary cells. *Biochem. J.* **313,** 991–996 (1996).
19. Rong, Y., Durden, D. L., Van Meir, E. G. & Brat, D. J. 'Pseudopalisading' necrosis in glioblastoma: a familiar morphologic feature that links vascular pathology, hypoxia, and angiogenesis. *J. Neuropathol. Exp. Neurol.* **65,** 529–539 (2006).
20. Nelson, D. A. et al. Hypoxia and defective apoptosis drive genomic instability and tumorigenesis. *Genes Dev.* **18,** 2095–2107 (2004).
21. Chaneton, B. et al. Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* **491,** 458–462 (2012).
22. Gui, D. Y., Lewis, C. A. & Vander Heiden, M. G. Allosteric regulation of PKM2 allows cellular adaptation to different physiological states. *Sci. Signal.* **6,** pe7 (2013).
23. Keller, K. E., Tan, I. S. & Lee, Y. S. SAICAR stimulates pyruvate kinase isoform M2 and promotes cancer cell survival in glucose-limited conditions. *Science* **338,** 1069–1072 (2012).
24. Anastasiou, D. et al. Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses. *Science* **334,** 1278–1283 (2011).
25. Christofk, H. R. et al. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452,** 230–233 (2008).
26. Anastasiou, D. et al. Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis. *Nature Chem. Biol.* **8,** 839–847 (2012).
27. Lunt, S. Y. & Vander Heiden, M. G. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annu. Rev. Cell Dev. Biol.* **27,** 441–464 (2011).
28. Papandreou, I., Cairns, R. A., Fontana, L., Lim, A. L. & Denko, N. C. HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metab.* **3,** 187–197 (2006).
29. Brown, J. M. & Wilson, W. R. Exploiting tumour hypoxia in cancer treatment. *Nature Rev. Cancer* **4,** 437–447 (2004).

METHODS

Materials. The following antibodies were used: antibodies to GLDC (HPA002318), SHMT2 (HPA020549) from Sigma; antibodies to actin (sc-1616), SHMT1 (sc-100849), and GCAT(sc-86466) from Santa Cruz; anti-GCSH (H00002653-A01) from Abnova; anti-SOX-2 (MAB2018) from R & D systems; anti-GFAP (IF03L) from Calbiochem; anti-cleaved-PARP (19F4) and anti-PKM2 (D78A4) from Cell Signaling Technologies; anti-GCAT (ab85202) from Abcam; anti-methylglyoxal antibody (MMG-030) from Genox; HRP-conjugated anti-mouse, anti-rabbit, and anti-goat secondary antibodies from Santa-Cruz Biotechnology.

The following cell culture reagents were used: neurobasal medium, N-2 and B-27 supplements from Invitrogen; recombinant human FGF basic (4114-TC) and EGF (236-EG) from R & D systems; DMEM and RPMI-1640 media, doxycycline (D9891) from Sigma; leucine ethyl ester hydrochloride (61850), arginine ethyl ester hydrochloride (A2883), alanine ethyl ester hydrochloride (855669), valine ethyl ester hydrochloride (220698), lysine ethyl ester dihydrochloride (62880), ethylamine (395064) from Sigma; glycine ethyl ester hydrochloride (sc-295020) and polybrene (sc-134220) from Santa Cruz.

Additional materials used: formalin from VWR; Borg Decloaker RTU solution and pressurized Decloaking Chamber from Biocare Medical; Prolong Gold Antifade reagent from Invitrogen; CellTiter-Glo Luminescent Assay from Promega; [$U^{14}C$] serine from MP Biomedicals; Matrigel (356230) from BD Biosciences.

Cell lines, tissue culture, and media. The neurosphere-forming lines 0308, BT145, and BT112 were established as described[10,30,31], provided by H. Fine and K. Ligon, and maintained as tumorigenic neural stem cell-like neurospheres in NBE medium (neurobasal medium containing N-2 and B-27 supplements, epidermal growth factor, basic fibroblast growth factor, L-glutamine, and penicillin-streptomycin) as described[10]. When passaging, spheres were manually broken into smaller spheres and single cells by trypsinization and pipetting. For differentiation experiments, neurospheres were broken into single cells and grown in DMEM (containing 10% inactivated fetal bovine serum and penicillin-streptomycin) for at least 1 week.

All other cell lines (LN229, ACHN, A2058, U251, T47D, MCF7, HMC-1-8, U87, DoTc2-4510, and PC3) were obtained from Broad Institute Cancer Cell Line Encyclopedia, and cultured as adherent cell lines in DMEM with exceptions noted below. Cell lines were verified to be free of mycoplasma contamination. Cell line origins are as follows: 0308, BT145, BT112, LN229, U251, U87 (glioblastoma), ACHN (renal cell adenocarcinoma), A2058 (melanoma), T47D (breast ductal carcinoma), MCF7 (breast pleural effusion), HMC-1-8 (breast pleural effusion), DoTc2-4510 (cervical carcinoma), PC3 (prostate adenocarcinoma). When comparing SHMT2 protein expression across cell lines, all cell lines were grown in NBE for 2 days before collection in order to be grown under identical conditions.

For experiments measuring oxygen consumption and for untargeted metabolite profiling experiments, RPMI was used, which does not contain the PKM2 product pyruvate.

Subcutaneous xenograft experiments. For regular subcutaneous xenograft studies, LN229 cells were transduced to stably express shRNAs (shGLDC^{dox}, shGFP via puromycin selection for 3 days) and then cDNAs (empty vector or RNAi-resistant *GLDC* cDNA via blasticidin selection for 3 days) then further amplified. Xenografts were initiated with 3 million cells injected subcutaneously per site, with 30% Matrigel, 100-μl injection volume in the left and right flanks of female, 6–8 week old NCr nude mice (Taconic). Tumours were allowed to form for two weeks, and at this point the first caliper measurements were taken, and induction started by addition of doxycycline at $2\,g\,l^{-1}$ to drinking water. Tumour volume was calculated using the modified ellipsoid formula ½(length × width^2) and expressed as relative fold change to the initial volume of each tumour at the start of doxycycline induction.

For rapid tumour xenograft studies to form ischaemic tumour cores, LN229 cells were transduced to stably express both cDNAs (empty vector, RNAi-resistant *SHMT2* cDNA, or *PKM2* cDNA via blasticidin selection) and shRNAs (via puromycin selection). Xenografts were initiated with 8 million cells injected subcutaneously per site in the left and right flanks of female, 6–8-week-old NCr nude mice (Taconic). Tumours were removed at 48 h post-injection and fixed in 10% formalin.

For the quantification of viable and nonviable regions in the ischaemic region, fixed tumours were embedded and sections prepared. Sections were immunostained for SHMT2 and cleaved PARP, and images of the central tumour regions were obtained using a Zeiss Axiovert 200M inverted fluorescent microscope and AxioVision Software. All images were acquired and processed under the same parameters across the entire set. The image labels were scrambled so that analyses could be carried out in a blinded manner, and the Red channel (cPARP) and Blue channel (Hoechst 33342) was analysed. Using Adobe Photoshop, the entire central necrotic region, labelled by cleaved PARP, was manually outlined with the Lasso tool. Within this tumour region, the total area counts (in pixels) of the dead (cPARP positive and Hoechst positive) and viable (cPARP negative and Hoechst positive) regions were obtained to calculate the percentage of viable region within the central necrotic zone.

Analyses of oncogenomic and other microarray data. We had previously classified a set of 2,752 metabolic enzymes and transporters[32]. To obtain a list of metabolic genes and transporters that have increased expression in gliomas, we analysed the 9 expression studies deposited in Oncomine[33] that profiled gene expression normal brain tissue and gliomas. For each data set, the top 10% of genes overexpressed in the glioma relative to normal brain was obtained, and cross referenced with our gene set, which yielded a list of 367 genes which placed within the top 10% of overexpressed genes in at least two separate studies.

To determine the expression of selected metabolic genes in the context of neural stem cells, we analysed a set of 5 microarray data sets deposited in Gene Expression Omnibus in which neural stem cells are compared with differentiated controls (GSE 36484, GSE10721, GSE15209, and two comparison groups in GSE11508). A summary of the data sets and fold change in expression of each gene in each study is provided in Supplementary Table 3.

shRNA expressing lentivirus generation and sequences. For each gene of interest (GLDC, SHMT2, GCAT, GCSH), 5 lentiviral shRNA constructs were obtained from The RNAi Consortium (TRC) and recombinant lentivirus containing supernatant was produced using a transient transfection protocol[34]. Each lentivirus was separately transduced into LN229 by overnight incubation of virus in trypsin dissociated cells (20,000 cells per ml, 2 ml into each well of a 6-well plate) in the presence of polybrene. Lentiviral expression of shGFP and shLacZ served as negative controls for gene knockdown, and noninfected cells served as negative controls for transduction. Cells were selected with puromycin for 3 days to ensure transduction, and for each gene, the two (or three) most effective shRNAs, in terms of knockdown of protein expression by western blot, were chosen for use in our experiments.

The following shRNA sequences were used: shGFP: TRCN0000072186, target sequence: TGCCCGACAACCACTACCTGA; shLacZ: TRCN000007 2235, target sequence: CCGTCATAGCGATAACGAGTT; shGLDC_1: TRCN0000036599, target sequence: CGAGCCTACTTAAACCAGAAA; shGLDC_2: TRCN0000036603, target sequence: GAAGTTTATGAGTCTCC ATTT; shGLDC^{dox}: target sequence same as shGLDC_2, cloned into doxycycline-inducible vector (pLKO_GC11); shSHMT2_1: TRCN0000238795, target sequence: CGGAGAGTTGTGGACTTTATA; shSHMT2_2: TRCN0000034804, target sequence: CCGGAGAGTTGTGGACTTTAT; shSHMT2_3: TRCN 0000234657, target sequence: GTCTGACGTCAAGCGGATATC; shGCSH_1: TRCN0000083395, target sequence: GTGAACTCTATTCTCCTTTAT; shG-CSH_2: TRCN0000428788, target sequence: TGAGGAACACCACTATC TTAA; shGCAT_1: TRCN0000034579, target sequence: CCTTAACTT CTGTGCCAACAA; shGCAT_2: TRCN0000034580, target sequence: CCAG AGGTTCCGTAGTAAGAT; shNOTCH2_1: TRCN0000004896, target sequence: CCAGGATGAATGATGGTACTA; shNOTCH2_2: TRCN0000 004897, target sequence: CCACACAACAACATGCAGGTT.

Cell viability assays with shRNA transduction. For cell viability experiments involving transduction of a single shRNA (for example, shGLDCs), cell lines (neurosphere-forming cell lines, LN229, ACHN, A2058, U251, T47D, MCF7, HMC-1-8, U87, DoTc2-4510, and PC3) were seeded in 96-wells at 3,500 to 5,000 cells per well. The next day, neurosphere-forming lines were infected with lentivirus and polybrene via 30-min spin at 2,250 r.p.m. followed by incubation for 1 h before a media change (due to neurosphere-forming cell line sensitivity to prolonged incubation with virus and polybrene), while all the non neurosphere cell lines were infected via overnight incubation of virus and polybrene before a media change. For all non-neurosphere cell lines, puromycin selection was started 24 h after infection, while for the neurosphere lines it was started 48 h after infection (because of their sensitive nature). Cells were incubated for 4–6 additional days as indicated, and overall cell viability was quantified using the Cell Titer Glo (CTG) reagent (Promega) and measuring luminescence. As doubling times and luminescence values per viable cell differ between different cell lines, values are normalized to the same cells transduced in parallel with innocuous shGFP hairpins as indicated.

When comparing sensitivity to GLDC (or GCSH, SHMT2) knockdown across different cell lines, two identical sets of experiments, one which receives puromycin selection and one which does not, were carried out in parallel. Comparing the two ensures that the toxicity observed in the 'sensitive' cell lines is due to GLDC knockdown and not due to selection of nontransduced cells, because identical toxicity is also seen in the nonselected plate. Conversely, we can ensure that low toxicity observed in the 'insensitive' cell lines is not an artefact of poor transduction because if they had been poorly transduced, then toxicity would be observed in the puromycin selected plate. In this manner, we verified full transduction of cells that we have examined for GLDC effects on viability.

For some experiments, cells are transduced with more than one shRNA and this was carried out in a sequential manner. Cells were infected with the first lentivirus expressing an shRNA (shGFP, shSHMT2_1, or shSHMT2_2, shGCAT_1, shGCAT_2) as described, then selected in puromycin for 3 days, and expanded for 2–5 more days. Equal numbers of each stable cell line were infected with the second lentivirus (shGFP,

shGLDC_1, or shGLDC_2), seeded in 96-well plates, and at 5 days following infection, cell viability was measured. Because in some cases (for example, shGCAT hairpins) the primary transduction itself moderately impairs cell proliferation, viability values for the cells secondarily transduced with shGLDCs are always expressed as relative to the same primary transduced cells, processed in parallel, which are secondarily infected with a control hairpin (shGFP). Because the secondary transduction cannot be selected for (since the cells are already puromycin resistant from the first round of transduction), effective knockdown of the second gene was verified by western blot.

For cell proliferation experiments, cell counts were determined using a Coulter counter (Beckman).

CRISPR-Cas9 mediated gene knockdown. In some of our experiments, effective gene knockdown was achieved via CRISPR/Cas-9 mediated genome editing. We used pLENTICRISPR, in which both single guide RNA, directed against a target of interest, and the Cas9 endonuclease are both delivered to cells via lentivirus[35] in an analogous manner to the TRC shRNA experiments. Three target site sequences, selected based on best scores as previously calculated for all genes[36], were cloned into pLENTICRISPR. As described for TRC shRNA transduction, lentiviruses were produced and transduced into trypsin dissociated LN229 cells, via overnight incubation with polybrene. Following media change and puromycin selection, cells were harvested 7 days following infection, and gene knockdown determined by western blotting. The two most effective target guide sequences, in terms of knockdown of protein expression by western blot, were chosen for use in our experiments.

The following target site sequences, transduced via pLENTICRISPR, were used:
sgGFP: TGAACCGCATCGAGCTGAAG (plus strand)
sgGLDC_1: CGGGACAGCAGCAGTGGCGG (minus strand)
sgGLDC_2: ATTTGGGGTAGACATCGCCC (minus strand)
sgGCAT_1: CCAGCGCTGACTGTGCGCGG (minus strand)
sgGCAT_2: GAAGCATCGGCTGCGCCTGG (plus strand)

Pooled shRNA screening. pLKO.1 lentiviral plasmids encoding shRNAs targeting glycine metabolizing enzymes, metabolic enzymes for other amino acids, as well as nontargeting controls were obtained and combined to form a pool as described in Extended Fig. 2. This pool was used to generate a pool of lentiviruses as described[34]. LN229 cells were infected with the pooled virus at a low titre (multiplicity of infection of 0.7) to ensure that each cell contained only one viral integrant. After cells were selected for 3 days with puromycin, pooled cells were dissociated, divided, and subjected to a secondary infection with either shGFP, shGLDC_1, or shGLDC_2. After 6 days following the secondary infection, a time point corresponding to moderate toxicity as determined by decreased proliferation and moderate changes cell morphology compared to shGFP infected cells, cells were collected to obtain genomic DNA. As previously described[32], the shRNAs encoded in the genomic DNA were amplified and analysed by high throughput sequencing (Illumina) using the following primers:

Barcoded forward primer ('N's indicate location of sample-specific barcode sequence): AATGATACGGCGACCACCGAGAAAGTATTTCGATTTCTTGG CTTTATATATCTTGTGGAANNGACGAAAC; Common reverse primer: CA AGCAGAAGACGGCATACGAGCTCTTCCGATCTTGTGGATGAATACTG CCATTTGTCTCGAGGTC. Illumina sequencing primer: AGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAA.

Sequencing reads were deconvoluted using GNU Octave software as described. For each shRNA, Abundance was defined as (number of reads/total number of all reads of pooled cells in either shGFP or shGLDC). Enrichment was defined as (abundance in shGLDC / abundance in shGFP), thus an enrichment score of 2.0 would indicate that an shRNA is twice as abundant in the shGLDC infected pool as it was in the shGFP infected pool. Fold change is defined as the enrichment score of an shRNA relative to the mean enrichment score of all 11 nontargeting control shRNAs. For a given gene, the mean fold change (shGLDC_1/shGFP and shGLDC_2/shGFP) was calculated from all shRNAs targeting that gene.

Clonogenic neurosphere formation assay. 0308 cells were stably transduced with shRNAs as indicated, and seeded at single cell-per-well density in poly-D-lysine coated 384-well plates (Becton Dickinson). Wells containing a single cell were marked and two weeks later, the marked wells containing spheres were counted.

Histology and immunohistochemistry. Immunohistochemical analyses were performed on discarded archival biopsy (7) and autopsy (7) specimens of glioblastoma, World Health Organization Grade IV, seen at the Departments of Pathology, Massachusetts General Hospital and NYU Langone Medical Center, from 2010 to 2013. Approval from respective Institutional Review Board was obtained, and because we used discarded tissue only, a waiver of informed consent was received. Formalin-fixed, paraffin-embedded brain biopsy tissues were stained with routine haematoxylin and eosin stain (H&E), and cases were reviewed by a neuropathologist (M.S.) to select the most representative block/s for immunohistochemical analysis. Paraffin sections of GBM tumours and normal brains, fixed in 10% formalin, were subjected to deparaffinization and antigen retrieval with Borg Decloaker RTU solution pressurized Decloaking Chamber (Biocare Medical). Antibodies were diluted in in 4% horse serum and 0.1% tween in PBS, which was also used for blocking. Vectastain ABC

immunoperoxidase detection kit (Vector Labs) and DAB+ substrate kit (Dako) was used for chromogenic labelling. It was noted that antigen presentation for SHMT2 was much weaker in autopsy sections compared to tumour biopsy sections, likely a result of post-mortem interval, and thus a more concentrated primary antibody incubation and longer chromogenic development was required for these sections to get comparable signal to the biopsy sections.

Images were acquired using an Olympus BX41microscope and CellSens[R] software. For immunofluorescence staining of GBM tumours and normal brains, as well as rapid tumour xenografts, fixed in 10% formalin, the same deparaffinization, antigen retrieval, and blocking/antibody incubation steps were used as above. Immunoreactivity was detected using Alexa-fluor 488 and 568 antibodies and nuclei labelled with Hoechst 33352 (Life Technologies), and Prolong[R] Gold antifade reagent (Life Technologies) was used as mounting medium. Images were acquired using a Zeiss Axiovert 200M inverted fluorescent microscope and AxioVision Software. For all image-based data, acquisition and processing steps were carried out using the same parameters across the entire set, aside from the increased antibody concentration and longer chromogenic development for the set of autopsy sections for SHMT2 immunostaining as described.

Amino acid analyses. Intracellular amino acids were extracted by hot water extraction, and proteins were removed with sulfosalicylic acid. The amino acids were separated by high-resolution ion-exchange chromatography and derivatized with ninhydrin, and analysed on a Hitachi L-8800 amino acid analyser[37]. Amino acids were normalized by wet pellet weight of the cells before extraction.

Quantitative CE–MS based metabolite profiling. Capillary electrophoresis mass spectrometry-based targeted quantitative analysis was performed on stably transduced LN229 cells, as previously described[38]. A total of 116 metabolites involved in glycolysis, pentose phosphate pathway, tricarboxylic acid (TCA) cycle, urea cycle, and polyamine, creatine, purine, glutathione, nicotinamide, choline, and amino acid metabolism were analysed and listed in Supplementary Table 5.

Metabolite extraction and LC–MS analysis. Untargeted metabolite profiling, flux experiments, and amino acetone measurements were performed on a Dionex UltiMate 3000 ultra-high performance liquid chromatography system coupled to a Q Exactive benchtop Orbitrap mass spectrometer, which was equipped with an Ion Max source and a HESI II probe (Thermo Fisher Scientific). External mass calibration was performed every 7 days.

For untargeted metabolite profiling and flux experiments, polar metabolites were extracted from cells growing in a 6-well dish using 400 μl of ice cold 80% methanol with 20 ng ml^{-1} valine-d8 as an internal extraction standard. After scraping the cells, 400 μl of chloroform was added before vortexing for 10 min at 4 °C, centrifugation for 10 min at 4 °C at 16,000g, and drying 150 μl of the upper methanol/water phase under nitrogen gas. Dried samples were stored at −80 °C then resuspended in 40 μl 50% acetonitrile/50% water immediately before analysis. Cells were usually left plated for 24–48 h after a media change before extraction in order to allow for media conditioning. Accordingly, U-[^{13}C]glucose labelling of cells was achieved by adding a concentrated stock to glucose-free RPMI media to a final concentration of 11.1 mM after 24 h of media conditioning. Chromatographic separation was achieved by injecting 10 μl of sample on a SeQuant ZIC-pHILIC Polymeric column (2.1 × 150 mm 5 μM, EMD Millipore). Flow rate was set to 100 μl per min, column compartment was set to 25 °C, and autosampler sample tray was set to 4 °C. Mobile Phase A consisted of 20 mM ammonium carbonate, 0.1% ammonium hydroxide. Mobile Phase B was 100% acetonitrile. The mobile phase gradient (%B) was as follows: 0 min 80%, 5 min 80%, 30 min 20%, 31 min 80%, 42 min 80%. All mobile phase was introduced into the ionization source set with the following parameters: sheath gas = 40, auxiliary gas = 15, sweep gas = 1, spray voltage = −3.1kV or +3.0kV, capillary temperature = 275 °C, S-lens RF level = 40, probe temperature = 350 °C. In experiments to measure steady-state levels, metabolites were monitored using a polarity-switching full-scan method. In experiments using U-[^{13}C]glucose tracing, metabolites were monitored using a targeted selected ion monitoring (tSIM) method in negative mode with the quadrupole centred on the M-H ion $m+1.5$, $m+2.5$, or $m+3.5$ mass with a 8 a.m.u. isolation window, depending on the number of carbons in the target metabolite. Resolution was set to 70,000, full-scan AGC target was set to 10^6 ions, and tSIM AGC target was set to 10^5 ions. For tracing experiments, samples were collected at various time points as indicated. Labelling rate was calculated from counts at 6 min, and detailed methods for determining the labelling rate and overall flux are provided in the first three worksheets of Supplementary Table 8. Data were acquired and analysed using Xcalibur v2.2 software (Thermo Fisher Scientific). Full-scan untargeted data was analysed using Progenesis CoMet v2.0 software (Nonlinear Dynamics) to identify differential peaks (Supplementary Table 4 and 5) and the identified metabolites with greatest predicted change were further analysed with Xcalibur. Retention times for selected metabolites appearing in the untargeted analyses (AICAR and SAICAR) were confirmed by running a standard. All standards were obtained commercially, except for SAICAR, which was synthesized enzymatically from AICAR and purified by ion-exchange chromatography as described[23].

Derivatization and LC–MS detection of aminoacetone. The protocol for aminoacetone derivatization with fluorenylmethyl chloroformate (FMOC-Cl) and subsequent detection via LC–MS was adapted from previous studies[39,40]. LN229 cells grown in 6-cm plates were quickly washed in cold PBS, then extracted with scraping in 500 μl acetonitrile containing 1 μM ethylamine as an internal control for sample recovery and derivatization efficiency. Following vortexing, centrifugation, and transfer of supernatant to eliminate insoluble material, potassium borate buffer at pH 10.4 (final concentration 33 μM) and FMOC-Cl (final concentration 400 μg ml^{-1}) were added. Samples were completely dried, 100 μl of water added followed by addition of 800 μl hexane. Following vortexing and centrifugation, the upper phase was transferred to a new tube, dried, and the pellet extracted in acetonitrile.

For LC separation, 10 μl of each biological sample was injected onto an Ascentis Express C18 2.1×150 mm (2.7-μm particle size) column (Sigma-Aldrich). Mobile phase A was 0.1% formic acid and mobile phase B was 0.1% formic acid in acetonitrile. The chromatographic gradient was as follows, all at a flow rate of 0.25 ml min^{-1}: 0–2 min: hold at 5% B; 2–20 min: increase linearly to 75% B; 20–20.1 min: increase linearly to 95% B; 20.1–24 min: hold at 95% B; 24–24.1 min: decrease linearly to 5% B; 24.1–28 min: hold at 5% B. The autosampler was held at 4 °C and the column compartment was held at 35 °C. To minimize carryover, blank injections were performed after every six analytical runs.

All mobile phase was introduced into the ionization source with the spray voltage set to $+3.0$ kV and the same temperature and gas parameter settings as described in the previous section. The MS data acquisition was performed by tSIM of aminoacetone-FMOC and ethylamine-FMOC (internal standard) with the resolution set at 70,000, the AGC target at 10^5, the maximum injection time at 150 ms, and the isolation window at $1.0\,m/z$. The full scan range was 150–2,000 m/z. Quantitation of the data was performed with Xcalibur v2.2 using a 5 p.p.m. mass tolerance by a researcher (E.F.) blinded to the identity of the samples.

Peak areas for aminoacetone-FMOC were normalized to peak areas for ethylamine-FMOC from the same sample, and further normalized to total protein (μg) and expressed relative to the control sample.

Oxygen consumption measurements. Oxygen consumption of LN229 cells was measured using an XF24 Extracellular Flux Analyzer (Seahorse Bioscience). 60,000 cells were plated per well the night before the experiments, and RPMI 8226 media (US Biological 9011) containing 2 mM glutamine and 10 mM glucose without serum was used as the assay media. Oxygen consumption measurements were normalized based on protein concentration obtained from the same plate used for the assay.

Lactate dehydrogenase (LDH)-linked pyruvate kinase activity assay. Concentrated (5–10 mg ml^{-1}) hypotonic lysate was prepared from cells by swelling on ice for 10 min in one equivalent of $1\times$ hypotonic lysis buffer (20 mM HEPES pH 7.0, 5 mM KCl, 1 mM MgCl$_2$, 2 mM DTT, 1 tablet in 10 ml Complete EDTA-free protease inhibitor (Roche)), then passing through a 26 gauge needle $3\times$, then spinning 10 min at 4 °C at 16,000g. Concentrated lysate was diluted 1:100 in $1\times$ hypotonic lysis buffer and immediately assayed with 500 μM final PEP, 600 μM final ATP, 180 μM final NADH, and 0.16 mg ml^{-1} LDH in $1\times$ reaction buffer (50 mM Tris pH 7.5, 50 mM KCl, 1 mM DTT) in 100 μl total. Decrease in NADH fluorescence was followed in a Tecan plate reader and a regression on the slope of the decrease was taken as the activity. Bradford assay was performed on the concentrated lysate and activities were normalized to total protein.

Mitochondrial isolation and glycine cleavage assay. Intact mitochondria were isolated from mechanically lysed cells using differential centrifugation as described[41], and the intact state of mitochondria verified using the JC-1 dye. Equal quantities of isolated mitochondria were resuspended in a buffer to support glycine cleavage activity as described[42] (100 mM KCl, 50 mM mannitol, 20 mM sucrose, 10 mM KH$_2$PO$_4$, 0.1 mM EGTA, 1 mM MgCl$_2$, 0.175 mM pyridoxal phosphate, 1 mM ADP, 25 mM HEPES, pH 7.4) with the addition of 1 μM NAD+, 2 μM tetrahydrofolate, and 10 μM beta-mercaptoethanol. Upon addition of U-^{14}C serine, the reaction mixture containing mitochondria were incubated for 40 min at 37 °C, and CO$_2$ produced by the reactions were collected in phenylethylamine-coated paper overnight at 30 °C, and the ^{14}C content was measured using a scintillation counter.

Statistics and animal models. All experiments reported in Figs 1–4 were repeated at least three times, except Figs 1f, 1k, 2b, and 4c, which were repeated twice; Figs 1c, 2a, 2f, and 4a were performed once. In addition, histological analyses experiments (Fig. 3c) and xenograft based experiments (Figs 1h–j, 3g, 4g) were performed once, with n's indicating the number of individual patient-based tumours or xenograft tumours. All centre values shown in graphs refer to the mean. t-tests were heteroscedastic to allow for unequal variance and distributions assumed to follow a Student's t distribution, and these assumptions were not contradicted by the data. No samples or animals were excluded from analysis, and sample size estimates were not used. Animals were randomly assigned to groups. Studies were not conducted blind with the exception of Fig. 3g and 4g. All experiments involving mice were carried out with approval from the Committee for Animal Care at MIT and under supervision of the Department of Comparative Medicine at MIT. No statistical methods were used to predetermine sample size.

30. Mehta, S. *et al.* The central nervous system-restricted transcription factor Olig2 opposes p53 responses to genotoxic damage in neural progenitors and malignant glioma. *Cancer Cell* **19**, 359–371 (2011).
31. Chudnovsky, Y. *et al.* ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state. *Cell Rep* **6**, 313–324 (2014).
32. Possemato, R. *et al.* Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* **476**, 346–350 (2011).
33. Rhodes, D. R. *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
34. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl Acad. Sci. USA* **105**, 20380–20385 (2008).
35. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
36. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
37. Ptolemy, A. S. *et al.* A 9-month-old boy with seizures and discrepant urine tryptophan concentrations. *Clin. Chem.* **57**, 545–548 (2011).
38. Kami, K. *et al.* Metabolomic profiling of lung and prostate tumor tissues by capillary electrophoresis time-of-flight mass spectrometry. *Metabolomics* **9**, 444–453 (2013).
39. Xiao, S. & Yu, P. H. A fluorometric high-performance liquid chromatography procedure for simultaneous determination of methylamine and aminoacetone in blood and tissues. *Anal. Biochem.* **384**, 20–26 (2009).
40. Kazachkov, M. & Yu, P. H. A novel HPLC procedure for detection and quantification of aminoacetone, a precursor of methylglyoxal, in biological samples. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **824**, 116–122 (2005).
41. Rotem, R. *et al.* Jasmonates: novel anticancer agents acting directly and selectively on human cancer cell mitochondria. *Cancer Res.* **65**, 1984–1993 (2005).
42. Jois, M., Hall, B., Fewer, K. & Brosnan, J. T. Regulation of hepatic glycine catabolism by glucagon. *J. Biol. Chem.* **264**, 3347–3351 (1989).
43. Brat, D. J. & Van Meir, E. G. Vaso-occlusive and prothrombotic mechanisms associated with tumor hypoxia, necrosis, and accelerated growth in glioblastoma. *Lab. Invest.* **84**, 397–405 (2004).

LETTER

Therapy–induced tumour secretomes promote resistance and tumour progression

Anna C. Obenauf[1], Yilong Zou[1,2]*, Andrew L. Ji[1]*, Sakari Vanharanta[1,3], Weiping Shu[1], Hubing Shi[4], Xiangju Kong[4], Marcus C. Bosenberg[5,6], Thomas Wiesner[7], Neal Rosen[8], Roger S. Lo[4] & Joan Massagué[1]

Drug resistance invariably limits the clinical efficacy of targeted therapy with kinase inhibitors against cancer[1,2]. Here we show that targeted therapy with BRAF, ALK or EGFR kinase inhibitors induces a complex network of secreted signals in drug-stressed human and mouse melanoma and human lung adenocarcinoma cells. This therapy-induced secretome stimulates the outgrowth, dissemination and metastasis of drug-resistant cancer cell clones and supports the survival of drug-sensitive cancer cells, contributing to incomplete tumour regression. The tumour-promoting secretome of melanoma cells treated with the kinase inhibitor vemurafenib is driven by down-regulation of the transcription factor FRA1. *In situ* transcriptome analysis of drug-resistant melanoma cells responding to the regressing tumour microenvironment revealed hyperactivation of several signalling pathways, most prominently the AKT pathway. Dual inhibition of RAF and the PI(3)K/AKT/mTOR intracellular signalling pathways blunted the outgrowth of the drug-resistant cell population in *BRAF* mutant human melanoma, suggesting this combination therapy as a strategy against tumour relapse. Thus, therapeutic inhibition of oncogenic drivers induces vast secretome changes in drug-sensitive cancer cells, paradoxically establishing a tumour microenvironment that supports the expansion of drug-resistant clones, but is susceptible to combination therapy.

Kinase inhibitors such as vemurafenib, erlotinib or crizotinib have shown clinical efficacy in melanoma with *BRAF* mutations, or in lung adenocarcinoma with *EGFR* mutations or *ALK* translocations, respectively[3–6]. Although complete responses are rare, the vast majority of patients show partial tumour regression or disease stabilization. However, drug resistance invariably develops, and most patients progress within 6–12 months[3–16], representing a common complication of targeted therapies that hampers long-term treatment success. The rapid emergence of clinical drug resistance may be facilitated by a small number of pre-existing cancer cells that are intrinsically resistant or poised to adapt to drug treatment quickly[17–19]. How these minority clones of drug-resistant cells react to the marked changes in the microenvironment during tumour regression is not known. A better understanding of this process could lead to treatments that improve the efficacy of current targeted anti-cancer drugs.

To model therapeutic targeting of heterogeneous tumour cell populations *in vivo*, we mixed a small percentage of vemurafenib-resistant A375 human melanoma cells (A375[R]), labelled with a TK-GFP-luciferase (TGL) vector, together with mostly non-labelled, vemurafenib-sensitive A375 cells, and injected the admixture (A375/A375[R], 99.95/0.05%) subcutaneously in mice (Extended Data Fig. 1a). After the tumours were established, we treated the mice with vemurafenib or vehicle, and monitored the growth of resistant cells by bioluminescent imaging (BLI) *in vivo* (Fig. 1a). Although vemurafenib treatment decreased the volume of sensitive tumours (A375 alone) (Extended Data Fig. 1b), the

number of admixed resistant cells in regressing tumours (A375/A375[R]) significantly increased compared to vehicle-treated controls (Fig. 1b). Green fluorescent protein (GFP) staining confirmed increased numbers of resistant cells in regressing tumours, and EdU or BrdU staining confirmed their increased proliferation rate compared to the vehicle-treated controls (Fig. 1c and Extended Data Fig. 1c, d). Tumours comprising only resistant cells showed no growth difference when treated with vehicle or vemurafenib (Fig. 1d), indicating that the growth advantage of resistant cells in regressing tumours was not caused by direct effects of vemurafenib on cancer or stromal cells.

Treatment of mixed A375 and A375[R] tumours with dabrafenib, another BRAF inhibitor (RAFi), or doxycycline-induced knockdown of *BRAF* had similar effects (Extended Data Fig. 1e–g). In line with these findings, A375[R] cells co-implanted with other vemurafenib-sensitive melanoma cell lines (Colo800, LOX and UACC62) also showed an up to eightfold growth increase compared to vehicle-treated control groups (Fig. 1e). Growth acceleration of the resistant population in a regressing tumour was also observed in the patient-derived[8] melanoma cell line M249 and its vemurafenib-resistant derivative M249[R4], driven by an *NRAS* mutation, a clinically relevant resistance mechanism (Fig. 1e and Extended Data Fig. 1h). In immunocompetent mice, vemurafenib treatment of tumours formed by melanoma cell lines derived from *Braf[V600E] Cdkn2a[−/−] Pten[−/−]* mice (YUMM1.1, YUMM1.7) also promoted growth of the admixed vemurafenib-resistant cells (YUMM1.7[R], B16) (Extended Data Fig. 1i, j).

Crizotinib- or erlotinib-treated mice containing tumours formed by *ALK*-driven (H3122) or *EGFR*-driven (HCC827) human lung adenocarcinoma cells, respectively, admixed with minority clones of intrinsically resistant cells from the same cell lineage (lung adenocarcinoma cells H2030 and PC9) or melanoma cells (A375[R]) also led to increased outgrowth of the resistant cells (Fig. 1e and Extended Data Fig. 1k–m). Local growth acceleration of resistant cells in the regressing subcutaneous tumours resulted in higher lung metastatic burden (Fig. 1f). Thus, drug-resistant cancer cells benefit from therapeutic targeting of surrounding drug-sensitive cells.

Circulating tumour cells can infiltrate and colonize tumours. This phenomenon, termed self-seeding[20], may contribute to the distribution of resistant clones to several metastatic sites. Mice implanted with sensitive A375 tumours were treated with vehicle or vemurafenib, and intracardially injected with TGL-labelled A375[R] cells (Fig. 1g). A375[R] cells were more efficiently attracted to vemurafenib-treated regressing tumours compared to vehicle-treated controls, with 95% (21 out of 22) and 12.5% (2 out of 16) efficiency, respectively, exhibiting substantial accumulation of resistant cells in regressing tumours by day 5 (Fig. 1g and Extended Data Fig. 1n). To evaluate the contribution of seeding by resistant circulating tumour cells to disease relapse, we intracardially injected resistant A375[R] cells or vehicle into tumour-bearing mice and compared

[1]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [2]Gerstner Sloan Kettering School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [3]MRC Cancer Unit, University of Cambridge, Cambridge CB2 0XZ, UK. [4]Division of Dermatology, Department of Medicine and Jonsson Comprehensive Cancer Center, University of California, Los Angeles, California 90095, USA. [5]Department of Pathology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [6]Department of Dermatology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [7]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [8]Molecular Pharmacology and Chemistry Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.
*These authors contributed equally to this work.
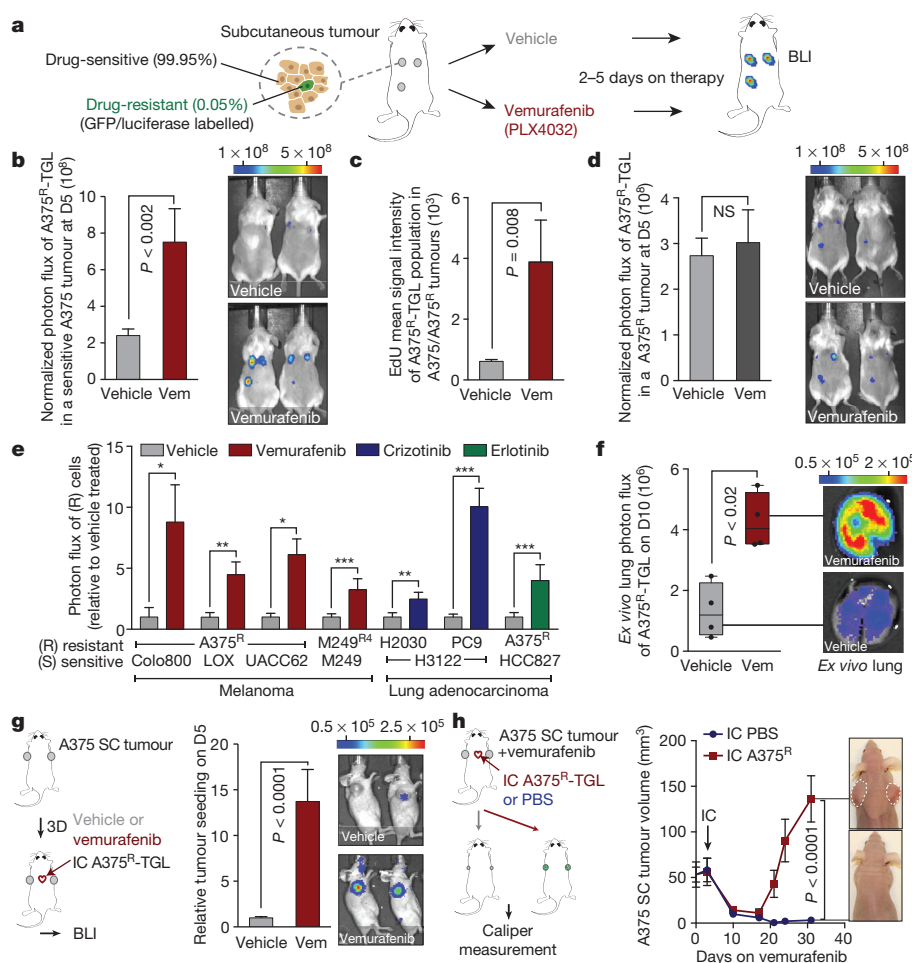
Figure 1 | The regressing tumour microenvironment stimulates the outgrowth, infiltration and metastasis of drug-resistant clones. a, Schematic of the experimental set-up. **b**, Bioluminescent signal of drug-resistant A375^R-TGL cells in vemurafenib-sensitive, A375 tumours, treated with vehicle or vemurafenib for 5 days (vehicle, $n = 36$; vemurafenib, $n = 15$ tumours). D, day. **c**, EdU incorporation in A375^R-TGL cells in A375/A375^R-TGL tumours treated with vehicle or vemurafenib for 4 days, as determined by FACS (vehicle, $n = 8$; vemurafenib, $n = 6$ tumours). **d**, Bioluminescent signal of A375^R-TGL tumours alone, treated with vehicle or vemurafenib for 5 days (vehicle, $n = 38$; vemurafenib, $n = 15$ tumours). **e**, Bioluminescent signal of TGL-expressing drug-resistant cancer cells (A375^R, M249^{R4}, PC9 and H2030) in drug-sensitive tumours (Colo800, LOX, UACC62, M249, H3122 and HCC827) treated with vehicle or drugs (vemurafenib, crizotinib and erlotinib) for 5 days (n (from left to right on the graph) = 6, 7, 12, 12, 9, 9, 25, 26, 9, 12, 12, 12, 16 and 11 tumours). **f**, Spontaneous lung metastasis by A375^R cells in mice bearing A375/A375^R-TGL tumours treated with vehicle or vemurafenib (10 days), visualized by BLI ($n = 4$). **g**, Seeding of A375^R-TGL cells from the circulation to unlabelled, subcutaneous (SC) A375 tumours of mice treated with vehicle or vemurafenib. Signal in the tumour was quantified by BLI (vehicle, $n = 30$; vemurafenib, $n = 34$ tumours; three independent experiments combined). IC, intracardiac. **h**, Treatment response, determined by tumour size, of subcutaneous A375 tumours allowed to be seeded by A375^R-TGL cells from the circulation or mock injected (vehicle, $n = 16$; vemurafenib, $n = 8$ tumours). Data in **b–e**, **g**, **h** are mean and s.e.m; in **f** the centre line is median, whiskers are minimum and maximum values. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-tailed Mann–Whitney U test. NS, not significant.

the tumour volume during vemurafenib treatment (Fig. 1h). Whereas the unseeded tumours in the control group showed extensive tumour regression, seeding by A375^R cells led to rapid tumour relapse (Fig. 1h). These results suggest that tumours regressing on targeted therapy are potent attractors of resistant circulating tumour cells that may contribute to rapid tumour progression.

Tumours consist of a complex microenvironment composed of immune, stromal and cancer cells[21]. Soluble mediators from this microenvironment can foster cancer growth and therapy resistance[13,14,22–24]. Considering that drug-sensitive cancer cells are the main population affected by targeted therapy, we proposed that signals derived from sensitive cancer cells in response to kinase inhibitors drive the outgrowth of drug-resistant cells. To test this hypothesis, we established an *in vitro* co-culture system and monitored the growth of TGL-expressing resistant cells (A375^R, H2030) in the absence or presence of sensitive cells treated with kinase inhibitors or vehicle (Fig. 2a). Mimicking our *in vivo* findings, co-culture with vemurafenib-, crizotinib- or erlotinib-treated sensitive cells significantly enhanced the growth of resistant cancer cells (Fig. 2a and Extended Data Fig. 2a–c).

We derived conditioned media (CM) from vemurafenib-sensitive melanoma cells cultured in the absence (CM-vehicle) or presence of vemurafenib (CM-vemurafenib). CM-vemurafenib accelerated the proliferation of drug-resistant cells, with different clinically relevant resistance mechanisms, as determined by cell viability assays and Ki67 staining (Fig. 2b and Extended Data Fig. 2d–f). Similarly, conditioned media from crizotinib- or erlotinib-treated sensitive lung adenocarcinoma cells stimulated proliferation of lung adenocarcinoma cells with intrinsic or

acquired resistance (Fig. 2c) and across different cell lineages (Extended Data Fig. 2g). In addition, CM-vemurafenib elicited increased cell migration in transwell migration and monolayer gap-closing assays (Fig. 2d and Extended Data Fig. 2h–k). CM-vemurafenib was also active on vemurafenib-sensitive cancer cells, increasing survival and suppressing the apoptotic caspase activity up to 100-fold in these cells when treated with vemurafenib *in vitro* (Fig. 2e, f). Because all biologically active conditioned media was collected before cell death or senescence, it is likely that the secretome is actively produced as a result of oncogene inhibition (Extended Data Fig. 2l, m). These results demonstrate that *BRAF*, *ALK* and *EGFR* mutant cells respond to therapeutic stress under targeted therapy by secreting factors that support the survival of drug-sensitive cells and accelerate the growth of drug-resistant minority clones. The effects of this reactive secretome may augment previously reported resistance mechanisms including relief of feedback inhibition of intracellular signalling[11,25], upregulation of receptor tyrosine kinases[26], or the supply of stromal cytokines[14] that protect the drug-sensitive cells.

To identify relevant components and regulators of the reactive secretome, we analysed gene expression changes in sensitive A375 melanoma cells at different time points after vemurafenib exposure *in vitro*. After 6 h on vemurafenib, 473 genes showed altered expression, and pathway analysis revealed that these genes were enriched for transcriptional regulators (Fig. 3a, b, Extended Data Fig. 3a, b and Supplementary Table 1). After 48 h, more than one-third of the transcriptome was differentially expressed (>5,000 genes; 405 genes encoding for proteins in the extracellular region, Gene Ontology (GO) accession 0005576), significantly overlapping with the gene expression changes of A375
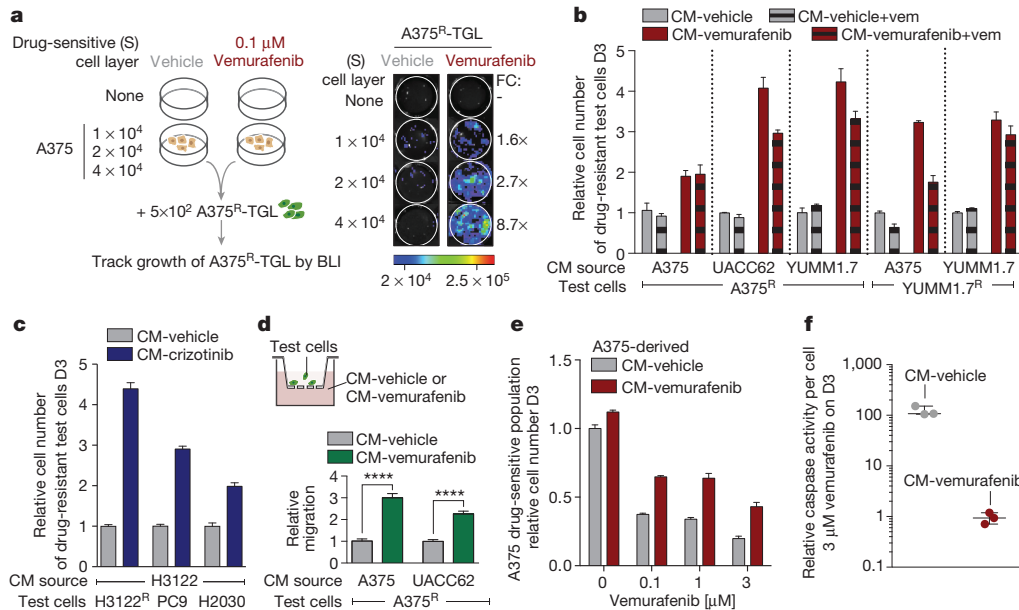
Figure 2 | The secretome of RAF and ALK inhibitor-treated tumour cells increases proliferation and migration of drug-resistant cells and supports the survival of drug-sensitive cells. a, Schematic (left) and representative BLI images (right) after 7 days of co-culture. Average fold change (FC) of BLI signal from A375R-TGL cells in vemurafenib-treated wells relative to vehicle-treated control wells is depicted on the right ($n = 4$ biological replicates). **b, c,** Conditioned media (CM) was derived from drug-sensitive cells, treated with vehicle, vemurafenib or crizotinib. Drug-resistant cells were grown in this conditioned media and the cell number was determined on day 3. Drug-sensitive and drug-resistant cell lines and drugs used to generate conditioned media as indicated. $n = 3$ (**b**) and 6 (**c**) biological replicates. **d,** Schematic diagram of the migration assay (top) and relative migration of A375R cells towards conditioned media from different sources as indicated (bottom, $n = 10$ fields of vision (FOV)). ****$P < 0.0001$, two-tailed Mann–Whitney U test. **e,** Survival assay of drug-sensitive A375 cells cultured in conditioned media and treated with vemurafenib, assessed on day 3 ($n = 3$ biological replicates). **f,** Apoptosis rate of A375 cells cultured in conditioned media and treated with vemurafenib (3 µM) ($n = 3$ biological replicates). Data are mean and s.e.m.
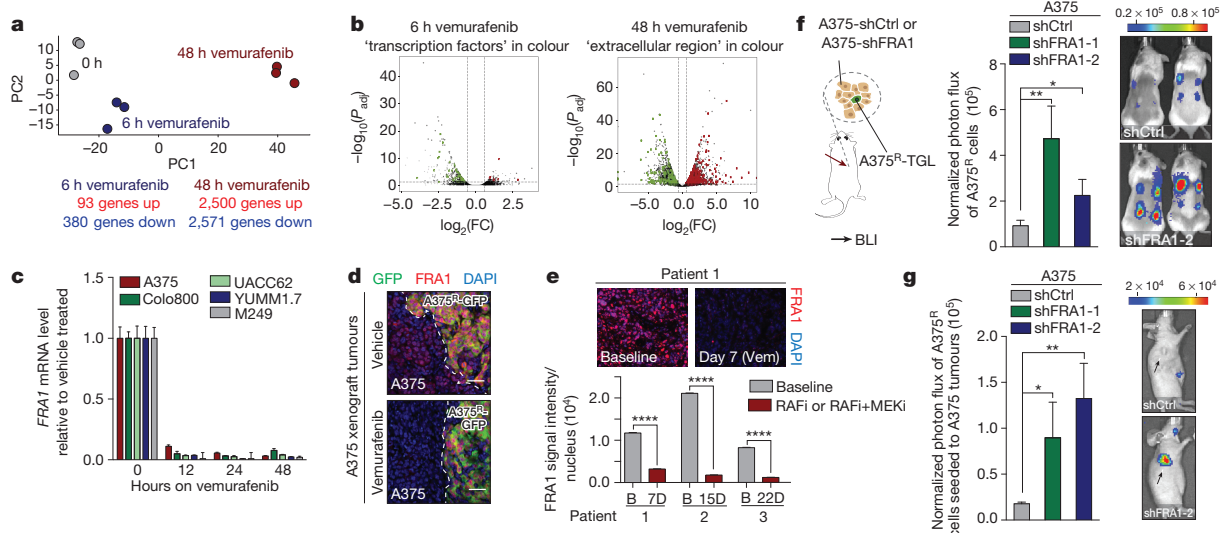
Figure 3 | FRA1 downregulation during RAFi treatment drives the reactive secretome. a, Principal component (PC) analysis of drug-sensitive A375 cells treated *in vitro* with vehicle or vemurafenib for 6 or 48 h. **b,** Volcano plots show genes significantly deregulated by vemurafenib treatment after 6 h (left) or 48 h (right). Transcription factors (TF) and gene products in the extracellular region are depicted in green (downregulated) and red (upregulated) ($n = 3$ tumours). P_{adj}, adjusted P value. **c,** Relative mRNA levels of *FRA1* during vemurafenib exposure [0.1–1 µM]. **d,** Representative immunofluorescence staining of A375/A375R tumours for GFP (A375R, green) and FRA1 (red) after vehicle or vemurafenib treatment (5 days). DAPI, 4′,6-diamidino-2-phenylindole. Scale bars, 50 µm. **e,** Top, representative immunofluorescence staining for FRA1 (red) of melanoma biopsy sections of patient 1. Original magnification, ×20. Bottom, nuclear FRA1 staining was quantified in three melanoma patients before (B) and early-on therapy. RAFi and MEKi denote RAF and MEK inhibitors, respectively. **f,** Bioluminescent signal of A375R-TGL cells 6 days after subcutaneous co-implantation with A375 cells expressing control (shCtrl) or two independent *FRA1* shRNAs (shFRA1-1 and shFRA1-2) ($n = 16$ tumours). **g,** Seeding of A375R-TGL cells to unlabelled tumours expressing control or two independent shRNAs for *FRA1*, determined by BLI (vehicle, $n = 10$; shFRA1-1, $n = 10$; shFRA1-2, $n = 8$ tumours). Data are mean and s.e.m. *$P < 0.05$, **$P < 0.01$, ****$P < 0.0001$, Student's t-test.

tumours *in vivo* after 5 days of vemurafenib treatment (Fig. 3a, b and Extended Data Fig. 3c). Similar extensive gene expression changes were observed in Colo800 and UACC62 melanoma cells treated with vemurafenib and H3122 lung adenocarcinoma cells treated with crizotinib (Extended Data Fig. 3d). Despite different cell lineages, different oncogenic drivers, and different targeted therapies we observed a significant overlap between the secretome of melanoma and lung adenocarcinoma cells ($P < 9.11 \times 10^{-5}$) (Extended Data Fig. 3e–h and Supplementary Table 1). Furthermore, changes in the secretome of vemurafenib-sensitive melanoma cells coincided with changes in the immune cell composition (Extended Data Fig. 4a, b), and with changes of soluble mediators derived from murine stromal cells such as IGF1 and HGF (Extended Data Fig. 4c, d). These data indicate a therapy-induced secretome (TIS), a response that consists of many up- and downregulated secreted factors, permeates the regressing tumour microenvironment and stimulates cancer cells, probably also stromal cells.

To identify molecular drivers of the A375-TIS in response to vemurafenib, we integrated the data of differentially expressed transcription factors after 6 h of vemurafenib treatment with the transcription factor binding motifs that were enriched at the promoters of differentially expressed genes in the secretome after 48 h (Fig. 3a, b). This analysis highlighted FRA1 (also known FOSL1), a member of the AP1 transcription factor complex and effector of the ERK pathway[27], as one of the putative upstream regulators of the TIS (Extended Data Fig. 5a). *FRA1* was downregulated in all drug-sensitive cells, but not in resistant cells, treated with vemurafenib, crizotinib and erlotinib (Fig. 3c, d and Extended Data Fig. 5b–d). Biopsies from melanoma patients early during RAFi treatment confirmed RAFi-induced FRA1 downregulation in clinical samples (Fig. 3e, Extended Data Fig. 5e and Extended Data Table 1).

To test the functional role of FRA1 in modulating the TIS, we used RNA interference (RNAi) to inhibit FRA1 expression. Co-culture and conditioned media assays using A375 cells expressing short hairpin

RNAs targeting *FRA1* (shFRA1) showed similar growth-accelerating and chemotactic activity on A375[R] cells as vemurafenib treatment (Extended Data Fig. 6a–d). In line with these results, FRA1 knockdown in A375 cells induced transcriptional changes similar to those induced by vemurafenib (Extended Data Fig. 6e). A375[R] cells co-implanted with A375 or UACC62 cells expressing shFRA1 also demonstrated increased growth *in vivo* (Fig. 3f and Extended Data Fig. 6f). A375-shFRA1 tumours attracted significantly more resistant cells from the circulation than tumours expressing the control vector (Fig. 3g). Thus, FRA1 downregulation drives the induction of the tumour-promoting secretome of vemurafenib-treated cancer cells.

To determine the effect of the reactive secretome on the drug-resistant tumour subpopulation in a regressing tumour, we expressed the ribosomal protein L10a (RPL10a) fused to enhanced green fluorescent protein (eGFP–RPL10a) in A375[R] cells, allowing the specific retrieval of transcripts from A375[R] cells by polysome immunoprecipitation for subsequent RNA-sequencing (RNA-seq) analysis[28] (Fig. 4a). In line with the *in vivo* phenotype of accelerated growth, the gene expression pattern of resistant cells in the regressing microenvironment was enriched for biological processes involved in cell viability, proliferation and cell movement (Extended Data Fig. 7a). Pathway analysis of the expression data suggested activation of several pathways including PI(3)K/AKT, BMP-SMAD and NFκB (Fig. 4b). The hyperactivity of the PI(3)K/AKT pathway in this context also suggested a potential vulnerability of the cells to PI(3)K/mTOR inhibitors (Extended Data Fig. 7b). The pathway-analysis-based prediction of PI(3)K/AKT activation was also reflected at the protein level in both resistant and sensitive cells in the presence of CM-vemurafenib *in vitro* and under vemurafenib treatment *in vivo* (Fig. 4c and Extended Data Fig. 7c, d). Moreover, PI(3)K/AKT emerged as the dominant TIS responsive pathway in a targeted immunoblot analysis of survival pathways *in vitro* (Extended Data Fig. 7e).
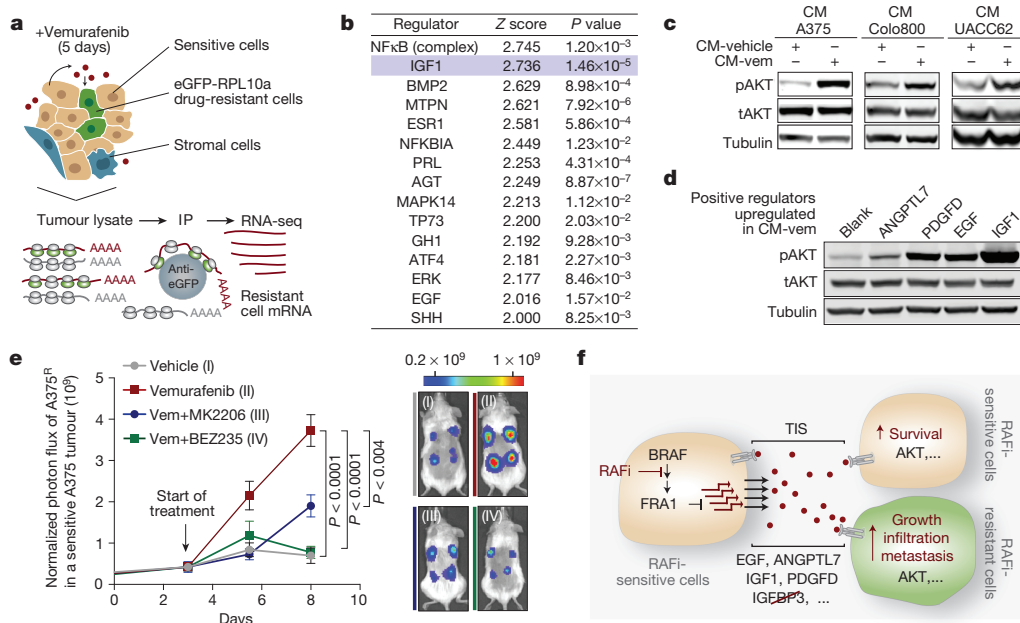
Figure 4 | The therapy-induced secretome in melanoma promotes relapse by activating the AKT pathway in resistant cells. a, Schematic diagram showing the isolation of polysome-associated transcripts from resistant cells by translating ribosome affinity profiling (TRAP) from tumours during treatment. IP, immunoprecipitation. **b**, Ingenuity upstream regulator analysis of gene expression profiles from A375[R] cells responding to a regressing tumour microenvironment (5 days of treatment; $n = 3$ tumours). **c**, Phosphorylation status of AKT[S473] (pAKT) in A375[R] cells, stimulated for 15 min with various conditioned media, as indicated by immunoblotting. tAKT, total AKT. **d**, Phosphorylation status of AKT[S473] in A375[R] cells after stimulation with

positive regulators of the AKT pathway, upregulated in the melanoma TIS; ANGPTL7 (5 µg ml^{-1}, 30 min; upregulated in A375, Colo800, UACC62), PDGFD (10 ng ml^{-1}, 10 min; upregulated in Colo800), EGF (10 ng ml^{-1}, 10 min; upregulated in A375) and IGF1 (10 ng ml^{-1}, 10 min; upregulated in UACC62). **e**, Mice bearing A375/A375[R]-TGL tumours were treated with drugs, and growth of A375[R] cells was followed by BLI (vehicle, $n = 14$; vemurafenib, $n = 16$; vemurafenib and BEZ235, $n = 16$; vemurafenib and MK2206, $n = 8$ tumours). **f**, Graphical summary of the findings. Data are mean and s.e.m. P values calculated using a two-tailed Mann–Whitney U test.

The TIS contained many mediators directly or indirectly activating the AKT pathway. Positive mediators that were upregulated during therapy included IGF1, EGF, ANGPTL7 and PDGFD, each of which activated the AKT pathway *in vitro* (Fig. 4d). IGF1, one of the most potent activators of the AKT pathway, is also abundantly expressed in the tumour stroma and is further upregulated during targeted therapy (Extended Data Figs 4c and 7f). In addition, levels of IGFBP3, a negative regulator of IGF1, were markedly reduced in the TIS of all investigated cell lines, favouring increased AKT pathway activation in the presence of IGF1 and stimulation of proliferation of resistant cells *in vivo* (Extended Data Fig. 7f–k).

To test the role of AKT activation as a mediator of TIS-induced tumour proliferation, we combined vemurafenib with AKT/PI(3)K/mTOR inhibitors. In co-culture and proliferation experiments using conditioned media, dual inhibition of the MAPK and AKT pathway diminished the growth benefit of the TIS (Extended Data Fig. 8a, b). We then treated mice with A375/A375^R or A375^R tumours with vemurafenib and AKT (MK2206) or PI(3)K/mTOR inhibitors (BEZ235). The combined inhibition of MAPK and PI(3)K/AKT/mTOR pathways significantly blunted the outgrowth of vemurafenib-resistant cells in the A375/A375^R tumours (Fig. 4e). The growth inhibition was specific for the amplified proliferation in the regressing tumour microenvironment and had no effects on the growth of resistant cells alone (Extended Data Fig. 8c). Furthermore, the outgrowth of resistant A375^R cells in tumour seeding assays was significantly reduced when regressing tumours were co-treated with BEZ235 (Extended Data Fig. 8d). Thus, the TIS-induced proliferation is susceptible to therapeutic targeting.

The limited effectiveness of targeted therapies has been attributed to intracellular feedback loops and specific cytokines that support the survival of drug-sensitive cells. From these residual tumours, clones emerge that are intrinsically resistant to targeted therapy and are ultimately responsible for clinical relapse. Our work demonstrates that targeted inhibition of a cancer driver pathway can paradoxically promote these two aspects of drug resistance via induction of a complex, reactive secretome. This TIS not only enhances the survival of drug-sensitive cells, but also acutely accelerates the expansion and dissemination of drug-resistant clones. Rather than a cell death by-product[29,30], the TIS is a live-cell response to inhibition of an oncogenic driver pathway, mediated by a concrete transcriptional program, and defined by specific alterations of intracellular signalling networks (Fig. 4f).

Our identification of AKT signalling as a mediator of TIS-induced tumour progression in BRAF-driven melanoma is in line with AKT activation in tumours observed in the clinic during vemurafenib treatment[16]. Patients treated with BRAF inhibitor rarely show full tumour regression[3,4], and the remaining drug-responsive tumour cells may remain a source of TIS for the duration of the treatment. Our results provide a rationale for combining PI(3)K/AKT/mTOR pathway inhibitors with inhibitors of the MAPK pathway in the treatment of these tumours. However, the breadth of the TIS and the generality of our findings across different cell lineages, drugs (vemurafenib, crizotinib and erlotinib), and resistance mechanisms suggest that durable responses may require the combination of this type of agents with a radically different therapeutic modality.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Engelman, J. A. & Settleman, J. Acquired resistance to tyrosine kinase inhibitors during cancer therapy. *Curr. Opin. Genet. Dev.* **18**, 73–79 (2008).
2. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Rev. Cancer* **13**, 714–726 (2013).
3. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
4. Sosman, J. A. *et al.* Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. *N. Engl. J. Med.* **366**, 707–714 (2012).
5. Shaw, A. T. & Engelman, J. A. ALK in lung cancer: past, present, and future. *J. Clin. Oncol.* **31**, 1105–1111 (2013).
6. Zhou, C. *et al.* Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol.* **12**, 735–742 (2011).
7. Villanueva, J. *et al.* Acquired resistance to BRAF inhibitors mediated by a RAF kinase switch in melanoma can be overcome by cotargeting MEK and IGF-1R/PI3K. *Cancer Cell* **18**, 683–695 (2010).
8. Nazarian, R. *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* **468**, 973–977 (2010).
9. Poulikakos, P. I. *et al.* RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature* **480**, 387–390 (2011).
10. Wagle, N. *et al.* Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *J. Clin. Oncol.* **29**, 3085–3096 (2011).
11. Lito, P. *et al.* Relief of profound feedback inhibition of mitogenic signaling by RAF inhibitors attenuates their activity in BRAFV600E melanomas. *Cancer Cell* **22**, 668–682 (2012).
12. Shi, H. *et al.* Melanoma whole-exome sequencing identifies V600EB-RAF amplification-mediated acquired B-RAF inhibitor resistance. *Nature Commun.* **3**, 724–728 (2012).
13. Wilson, T. R. *et al.* Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors. *Nature* **487**, 505–509 (2012).
14. Straussman, R. *et al.* Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. *Nature* **487**, 500–504 (2012).
15. Johannessen, C. M. *et al.* A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature* **504**, 138–142 (2013).
16. Shi, H. *et al.* A novel AKT1 mutant amplifies an adaptive melanoma response to BRAF inhibition. *Cancer Discovery* **4**, 69–79 (2014).
17. Diaz, L. A. Jr *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).
18. Shi, H. *et al.* Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discovery* **4**, 80–93 (2014).
19. Sharma, S. V. *et al.* A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).
20. Kim, M. Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315–1326 (2009).
21. Villanueva, J. & Herlyn, M. Melanoma and the tumor microenvironment. *Curr. Oncol. Rep.* **10**, 439–446 (2008).
22. Acharyya, S. *et al.* A CXCL1 paracrine network links cancer chemoresistance and metastasis. *Cell* **150**, 165–178 (2012).
23. Sun, Y. *et al.* Treatment-induced damage to the tumor microenvironment promotes prostate cancer therapy resistance through WNT16B. *Nature Med.* **18**, 1359–1368 (2012).
24. Lee, H. J. *et al.* Drug resistance via feedback activation of Stat3 in oncogene-addicted cancer cells. *Cancer Cell* **26**, 207–221 (2014).
25. Lito, P., Rosen, N. & Solit, D. B. Tumor adaptation and resistance to RAF inhibitors. *Nature Med.* **19**, 1401–1409 (2013).
26. Sun, C. *et al.* Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. *Nature* **508**, 118–122 (2014).
27. Joseph, E. W. *et al.* The RAF inhibitor PLX4032 inhibits ERK signaling and tumor cell proliferation in a V600E BRAF-selective manner. *Proc. Natl Acad. Sci. USA* **107**, 14903–14908 (2010).
28. Heiman, M. *et al.* A translational profiling approach for the molecular characterization of CNS cell types. *Cell* **135**, 738–748 (2008).
29. Fuchs, Y. & Steller, H. Programmed cell death in animal development and disease. *Cell* **147**, 742–758 (2011).
30. Kurtova, A. V. *et al.* Blocking PGE$_2$-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature* **517**, 209–213 (2015).

Supplementary Information is available in the online version of the paper.

Author contributions A.C.O. and J.M. conceived the project, designed the experiments and wrote the paper. A.C.O. performed experiments and computational analysis. A.L.J., Y.Z., W.S. and T.W. assisted with experiments. Y.Z. and S.V. performed computational analysis. M.C.B. provided cell lines. X.K., H.S. and R.S.L. provided patient samples. N.R. provided clinical expertise, cell lines and drugs. All authors interpreted data, discussed results, and revised the manuscript.

Author Information All RNA-seq data has been deposited in the Gene Expression Omnibus database under accession number GSE64741. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M. (j-massague@ski.mskcc.org).

METHODS

Cell culture. A375, M249 (ref. 8) and B16 cells were cultured in DMEM media; Colo800, UACC62, SKMEL239-clone3, LOX, PC9, H2030, H3122 and HCC827 cells were cultured in RPMI media. YUMM1.1 and YUMM1.7 were cultured in DMEM/F12 media. GPG29 and 293T cells were used for retrovirus and lentivirus production, respectively. Both were maintained in DMEM media. All media contained 10% FBS, 2 mM L-glutamine, 100 IU ml^{-1} penicillin/streptomycin and 1 μg ml^{-1} amphotericin B, the media for GPG29 contained in addition 0.3 mg ml^{-1} G418, 20 ng ml^{-1} doxycycline and 2 μg ml^{-1} puromycin. All cells were grown in a humidified incubator at 37 °C with 5% CO_2 and were tested regularly for mycoplasma contamination. All cell lines used were negative for mycoplasma.

To generate vemurafenib-resistant melanoma cell lines, vemurafenib-sensitive cell lines were seeded at low density and exposed to 1–3 μM vemurafenib (LC-Labs). After approximately 8 weeks of continuous vemurafenib exposure, we derived resistant cell clones that were maintained on vemurafenib (1 μM vemurafenib for M249^{R4}, Colo800^R, LOX^R, UACC62^R; 2 μM vemurafenib for A375^R, YUMM1.7^R). The same protocol was performed to generate a crizotinib-resistant cell line from H3122 lung adenocarcinoma cells, which were selected and maintained with 300 nM crizotinib. Drug-sensitive and resistant melanoma cell lines from A375, Colo800, UACC62 and YUMM1.7 and the drug sensitive lung adenocarcionoma cell lines H3122 and HCC827 were exposed to increasing doses of vemurafenib and the number of cells was determined after 3 days and pERK levels after 1 h of vemurafenib, crizotinib or erlotinib exposure (Extended Data Fig. 9a–j). Receptor status was determined by western blot and showed an increase in EGFR expression levels in all resistant lines examined as well as an increase in MET receptor expression in A375^R and UACC62^R cells compared to their parental, drug-sensitive cells (Extended Data Fig. 9k).

For co-culture assays sensitive cells were plated in 12-well or 24-well plates and allowed to adhere overnight in regular growth media. Media was then replaced with low serum (2% FBS) media containing vehicle, 0.1 μM vemurafenib, 0.3 μM crizotinib or 0.01 μM erlotinib. For control wells media containing vehicle or 0.1 μM vemurafenib, 0.3 μM crizotinib, or 0.01 μM erlotinib was plated at the same time. After 48 h, TGL-expressing, resistant cells were plated on top of the vehicle/drug treated cells or in media-only control wells. Media containing vehicle/drug was replenished every 48 h. After 7 days, luciferin [150 μg ml^{-1}] was added to the wells and luciferase-signal of resistant cells was determined by BLI using a Xenogen Spectrum imaging machine (Perkin Elmer). Co-culture experiments were independently performed at least twice and a representative experiment is shown.

To generate conditioned media, 2.3×10^6 and 6.4×10^6 drug-sensitive cells were plated on 15-cm dishes in regular growth media and allowed to adhere overnight. The media was then replaced by low serum media containing vehicle or vemurafenib (0.1 μM for A375 cells, 1 μM for all other cell lines), on dishes containing 2.3×10^6 and 6.4×10^6 drug-sensitive cells, respectively. The same procedure was followed for generation of conditioned media from H3122 (crizotinib, 0.3 or 1 μM) or HCC827 (erlotinib 0.01 μM) lung adenocarcinoma cells. After 72 h, cells on both plates had reached equal confluency of ~80% and conditioned media was collected, centrifuged at 1,000 r.p.m. for 5 min, filtered, and aliquots were stored at −80 °C until further use. Key proliferation and migration experiments yielded the same results when performed with conditioned media in which the same number of drug-sensitive cells (3.2×10^6) was plated initially, which resulted in higher cell confluency in the vehicle-treated dish at time of conditioned media collection.

Proliferation, survival and apoptosis assays. Around 1,000–3000 cells were plated in a 96-well plate, allowed to adhere overnight, and then incubated with either fresh or conditioned media containing vemurafenib or additional drugs as indicated. After 72 h, the number of cells was determined using a CelltiterGlo assay and the caspase 3/7 activity using a CaspaseGlo assay (Promega) according to the manufacturer's instructions. Caspase 3/7 activity was normalized to the number of cells present. All experiments with melanoma test cells and melanoma conditioned media were performed at least three times, experiments with lung adenocarcinoma cell lines were performed at least twice. Representative experiments are shown.

Boyden chamber transwell migration assay/gap closure assay. Transwell migration assays were performed as described previously with minor modifications[31]. In brief, serum-starved cells (0.2% FBS, overnight) were labelled with cell tracker green (Invitrogen) for 30 min at 37 °C and allowed to recover for 1 h. Cells (25,000–50,000) were then seeded onto membrane inserts with 8-μm pores and fluorescence blocking filters (Falcon). The number of cells migrated through the pores of the membrane was scored after 5–24 h using an Evos microscope (AMG). Gap closing assay was performed according to standard protocols. In brief, cells were seeded and grown until confluent. A tip was used to generate a gap, cells were washed and conditioned media was added. Images were acquired over time to monitor for cap closure in different conditions. All experiments were performed independently at least twice. Representative experiments are shown.

xCELLigence migration assay. Experiments were performed using the xCELLigence RTCA DP instrument (Roche Diagnostics GmbH) placed in a humidified incubator at 37 °C with 5% CO_2. Cell migration experiments were performed using modified 16-well plates (CIM-16, Roche Diagnostics GmbH) according to the manufacturer's instructions. The experiment was performed twice. A representative experiment is shown.

Animal studies. All experiments using animals were performed in accordance to our protocol approved by MSKCC's Institutional Animal Care and Use Committee (IACUC). 5–7-week-old, female NOD-SCID NCR (NCI) or athymic NCR-NU-NU (NCI) mice were used for animal experiments with human cell lines. Primary YUMM1.1 and YUMM1.7 cell lines were isolated from melanomas developed in mice (*Tyr::CreER; Braf^{CA}; Cdkn2a$^{-/-}$ Pten$^{lox/lox}$*) treated with 4-hydroxytamoxifen and were subsequently implanted in female C57BL/6J (JAX) mice aged between 5 and 7 weeks. Tumour formation, outgrowth and metastasis were monitored by BLI of TGL-labelled tumour cells as described previously[22]. In brief, anaesthetized mice (150 mg kg^{-1} ketamine, 15 mg kg^{-1} xylazine or isoflurane) were injected retroorbitally with D-luciferin (150 mg kg^{-1}) and imaged with an IVIS Spectrum Xenogen machine (Caliper Life Sciences). Bioluminescence analysis was performed using Living Image software, version 4.4. For co-implantation assays, mice were anaesthetized (150 mg kg^{-1} ketamine, 15 mg kg^{-1} xylazine) and 1×10^3 TGL-labelled resistant tumour cells were injected subcutaneously with 2×10^6 sensitive tumour cells in 50 μl growth-factor-reduced Matrigel/PBS (1:1) (BD Biosciences). For the control groups in which the effects of drug treatment on resistant cells alone were tested, 2×10^6 resistant cells were injected in growth-factor-reduced Matrigel/PBS. Two-to-four sites on the flanks were injected per mouse. After tumours reached a size of 50–150 mm^3, the BLI signal of resistant cells was determined. To compensate for minor growth differences of the GFP$^+$ resistant cell population between mice, the mice were assigned to the cohorts so that the overall BLI intensity (and consequently the cell number) was equal in the treatment and control group. Each group received vehicle or drug treatment as indicated (vemurafenib/PLX4032, 25 mg kg^{-1} twice daily for YUMM1.1 and YUMM1.7 tumours, and 75 mg kg^{-1} twice daily for all other BRAF mutant tumours, LC-Labs or Selleckchem; 100 mg kg^{-1} crizotinib once daily, LC-Labs; 50 mg kg^{-1} erlotinib once daily, LC-Labs; 100 mg kg^{-1} MK-2206 once daily, Chemietek; 50 mg kg^{-1} BEZ235 once daily, LC-Labs). Growth of the resistant population in the different groups was monitored by BLI, quantified and normalized to BLI signal at start of treatment. Tumour seeding and metastasis assays were performed as described with minor modifications[20]. In brief, sensitive tumour cells were injected subcutaneously on two sites per mouse. Once tumours were established (50–150 mm^3) mice were treated with vehicle or vemurafenib (75 mg kg^{-1} twice daily) for 3 days, and 1×10^5 TGL-labelled drug-resistant cells were injected in the left cardiac ventricle. Treatment was continued, and metastatic burden and tumour seeding were determined *in vivo* and *ex vivo* by BLI. Tumour volume was determined using caliper measurements and calculated using the following formula: tumour volume $= (D \times d^2)/2$, in which D and d refer to the long and short tumour diameter, respectively. All experiments with A375 cells were independently performed at least three times, except animal experiments in Fig. 3, which were performed twice. All other animal experiments were independently performed at least twice. Representative experiments are shown, except where noted and where instead the average of three experiments is presented.

Gene expression analysis. Whole RNA was isolated from cells using RNAeasy Mini Kit (QIAGEN). The Transcriptor First Strand cDNA synthesis kit (Roche) was used to generate cDNA. Differential RNA levels were assessed using Taqman gene expression assays (Life technologies). Assays used for human genes are: Hs04187685, Hs00365742, Hs00605382, Hs00601975, Hs01099999, Hs00959010, Hs01029057, Hs00234244, Hs00905117, Hs00180842, Hs00989373, Hs00234140, Hs00195591, Hs00207691, Hs99999141, Hs01117294, Mm00607939, Mm99999915 and Mm04207958. Relative gene expression was normalized to internal control genes: *B2M* (Hs99999907_m1), *GAPDH* (Hs99999905_m1) and *ACTB* (Mm00607939_s1). Quantitative PCR reactions were performed on a VIIA7 Real-Time PCR system and analysed using VIIA7 software (Life Technologies). All data points represent at least four technical replicates and experiments were performed independently three times. A representative experiment is shown.

Cancer-cell-specific TRAP and sequencing. To investigate the gene expression changes specifically of drug-sensitive tumours during vemurafenib treatment, or gene expression changes of resistant cells exposed to a regressing tumour microenvironment, A375 and A375^R cells, respectively, were modified to express eGFP-RPL10a. Tumours derived from implanted A375-eGFP-RPL10a and A375^R-eGFP-RPL10a cells were homogenized and processed with the TRAP protocol as previously described[28,32,33] with the following modifications: fresh tumour was homogenized with a Model PRO 200 homogenizer at speed 5 for four cycles of 15 s, RNasin Plus RNase inhibitor (Promega, N2615) was used as RNase inhibitor, and anti-eGFP antibody coated sepharose beads (GE Healthcare) were used for immunoprecipitation. Polysome-associated RNA was purified with RNAqueous micro kit (Life

Technologies, AM1931). Ribogreen and the Agilent BioAnalyzer technologies were used to quantify and control the quality of RNA; 500 ng RNA (RNA integrity number (RIN) > 8.5) from each sample was used for library construction with TruSeq RNA Sample Prep Kit v2 (Illumina) according to the manufacturer's instructions. The samples were barcoded and run on a Hiseq 2000 platform in a 50-base-pair (bp)/50-bp or 75-bp/75-bp paired-end run, using the TruSeq SBS Kit v3 (Illumina). An average of 40 million paired reads was generated per sample.

RNA-seq analysis. For drug-sensitive A375, Colo800, UACC63 and H3122 cells, *in vitro*, raw paired-end sequencing reads were mapped to the human genome (build hg19) with STAR2.3.0e (ref. 34) using standard options. Uniquely mapped reads were counted for each gene using HTSeq v0.5.4 (ref. 35) with default settings. Read counts of each sample were normalized by library size using the 'DESeq'[35] package of Bioconductor. Differential gene expression analysis between any two conditions was performed based on a model using the negative binomial distribution[35]. Genes with false discovery rate (FDR) < 0.05, fold change larger than 1.5 or smaller than 0.667-fold, and average read counts larger than 10 were treated as differentially expressed genes. RNA-seq data from *in vivo* xenograft TRAP samples were processed with the following modifications to avoid potential mRNA contamination from host mouse tissue: raw sequencing reads were mapped to a hybrid genome consisting indexes of both human (build hg19) and mouse (build mm9) genomes. Only reads that uniquely mapped to human genome indexes were preserved and counted using HTSeq v0.5.4 (ref. 35).

Bioinformatics analysis. Heatmap visualization of data matrices was performed using the 'gplots' package of R. Principle component analysis of RNA-seq results was performed with the variance stabilizing transformation methods in 'DESeq' package of Bioconductor and the first two principal components were plotted. Volcano plots were derived from 'DESeq'-based differential gene expression analysis. Differentially expressed genes with transcription factor activity (GO:00037000) at 6 h of vemurafenib treatment and gene products located in the extracellular region (GO:00005576) at 48 h of vemurafenib treatment were identified using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[36] v6.7 (http://david.abcc.ncifcrf.gov/) and enriched GO terms were visualized using REVIGO[37] (http://revigo.irb.hr). Enriched transcriptional regulators for the list of differentially expressed gene products in the extracellular region were predicted with DAVID v6.7 and this list compared to the gene expression levels of transcription factors after 6 h of vemurafenib treatment in A375 cells. Upstream regulators, functions associated with the gene expression profile and potential drug vulnerabilities were determined by interpretative phenomenological analysis (IPA) analysis on differentially expressed genes from A375[R]-eGFP-RPL10a cells in different tumour microenvironments *in vivo*.

Immunoblotting. RIPA buffer (Cell Signaling) was used for cell lysis, according to the manufacturer's instructions, and the protein concentrations were determined by BCA Protein Assay kit (Pierce). Proteins were separated by SDS–PAGE using Bis-Tris 4–12% gradient polyacrylamide gels in the MOPS buffer system (Invitrogen) and transferred to nitrocellulose membranes (BioRad) according to standard protocols. Membranes were immunoblotted with antibodies against pERK[T202/Y204] (4370), tERK (4696), pAKT[S473] (4060), pAKT[T308] (4056), tAKT (2920), EGFR (4267), MET (8198), PDGFRb (3169), pFRA1 (3880), caspase3 (9662), pPRAS40[T246] (13175), p70S6K[T389] (9205), pFAK[Y397] (3283), pPKC[betaIIS660] (9371), pNFκB[S536] (3033), pβ-Catenin[S33/37/T41] (9561), pSTAT-3[Y705] (9145), pSTAT-5[C11C5] (9359), pGSK3α/β[S21/9] (9327), pCREB[S133]/pATF-1 (9196) (Cell Signaling, 1:1,000), FRA1 (sc-605, Santa Cruz Biotechnology, 1:200) and tubulin (T6074, Sigma-Aldrich, 1:5,000) in Odyssey blocking buffer (LI-COR). After primary antibody incubation, membranes were probed with IRDye 800CW donkey-anti-mouse IgG (LI-COR) or IRDye 680RD goat-anti-rabbit IgG (LI-COR) secondary antibody (1:20,000) and imaged using the LI-COR Odyssey system. All immunoblots were performed independently at least twice. Tubulin served as a loading control.

Plasmids, recombinant protein and ELISA. Identifiers for shRNAs used in this study are: V3LHS-644610 (shFRA1-1), V3LHS-644611 (shFRA1-2), V3LHS-320021 (shIGFBP3-1) and V2LHS-111629 (shIGFBP3-2) (Dharmacon, GE Lifesciences). IGFBP3 ELISA (Raybiotech) was performed according to the manufacturer's instructions with 50 μg tumour lysate and conditioned media was diluted 1:5. Recombinant proteins were used at the following conditions: 10 ng ml^{-1} IGF1 (Invitrogen), 10 ng ml^{-1} EGF (Invitrogen), 10 ng ml^{-1} PDGFD (R&D Systems), 2 μg ml^{-1} IGFBP3 (Prospec) for 15 min, or 5 μg ml^{-1} ANGPTL7 (R&D Systems) for 30 min.

Patient samples. Melanoma tissues were obtained from clinical trial patients or patients under standard clinical management with approval of the UCLA Institutional Review Board. Patient-informed consent was obtained for the research performed in this study.

Immunofluorescence. Tissues for BrdU-immunofluorescence staining were obtained after overnight fixation with 4% paraformaldehyde (PFA) at 4 °C, embedded in OCT compound (VWR) and stored at −80 °C. 10-μm thick cryosections on glass slides were used for immunofluorescence staining according to standard

protocols. Tissue for all other immunofluorescence experiments from xenograft tumours was obtained after fixation with 4% PFA at 4 °C and a series of dehydration steps from 15% to 30% sucrose, as described previously[38]. In brief, tumours were sliced using a sliding microtome (Fisher). Tumour slices (80 μm) were blocked floating in 10% NGS, 2% BSA, 0.25% Triton in PBS for 2 h at room temperature. Primary antibodies were incubated overnight at 4 °C in the blocking solution and the next day for 30 min at room temperature. After washes in PBS-Triton 0.25%, secondary antibodies were added in the blocking solution and incubated for 2 h. After extensive washing in PBS-Triton 0.25%, nuclei were stained with Bis-Benzamide for 5 min at room temperature, tumour slices were washed and transferred to glass slides. Slices were mounted with ProLong Gold anti fade reagent (Invitrogen). Primary antibodies: GFP (GFP-1020, Aves Labs, 1:1,000), collagen IV (AP756, Millipore, 1:500), BrdU (ab6326, Abcam, 1:250), FRA1 (sc605, Santa Cruz, 1:200). Secondary antibodies: Alexa-Fluor-488 anti-chicken, Alexa-Fluor-555 anti-rabbit, Alexa-Fluor-555 anti-rat (Invitrogen). Stained sections were visualized using a Carl Zeiss Axioimager Z1 microscope or with a Leica SP5 upright confocal microscope using ×10 or ×20 objectives. Images were analysed with ImageJ, and Metamorph software.

Flow cytometry. Flow cytometry was performed as described previously[22], with minor modifications. In brief, whole tumours were dissected, cut into smaller sections and dissociated for 1–3 h with 0.5% collagenase type III (Worthington Biochemical) and 1% dispase II (Roche) in PBS. Resulting single cells suspensions were washed with PBS supplemented with 2% FBS and filtered through a 70-μm nylon mesh. The resulting single cell suspension was incubated for 10 min at 4 °C with anti-mouse Fc-block CD16/32 antibody (2.4G2 BD) in PBS supplemented with 1% BSA. Cells were subsequently washed with PBS/BSA and stained with control antibodies or antibodies to detect immune cells diluted in PBS supplemented with 0.5% BSA and 2 mM EDTA. The following antibodies against mouse antigens were used: CD45-PE-Cy7 (clone 30-F11, BD Pharmingen, 1:200), CD11b-APC (clone: M1/70, BD Pharmingen, 1:100), Gr1-PE (MACS, 1:10), CD31-APC (clone: 390, eBioscience, 1:100), F4/80-PE (clone: BM8, eBioscience, 1:50). To determine the level of EdU incorporation in A375[R] cells within vehicle- or vemurafenib-treated A375/A375[R] tumours, EdU (50 mg kg^{-1}, Life Technologies) was injected intraperitoneally, after 2 h tumours were collected, single-cell suspensions generated as described above and further processed according to the manufacturer's protocol (Click-iT Plus EdU Alexa Fluor 647 Flow Cytometry Assay Kit, Life technologies). Data were acquired using a FACS Calibur (BD Biosciences). All experiments were performed independently at least two times. Representative experiments are shown.

Antibody arrays. Cytokines and cytokine receptors of murine stromal and immune cells, in A375 tumours treated with vehicle or vemurafenib for 5 days, were measured using the Mouse Cytokine Array G2000 (RayBio, AAH-CYT-G2000-8, detecting 174 proteins), according to the recommended protocols. In brief, tumours were homogenized with a Mini Immersion Blender (Pro Scientific) in Raybio Lysis buffer with protease inhibitors. Lysates were centrifuged for 5 min at 10,000g, the supernatant was collected and protein concentration was measured using the BCA Assay Kit (Pierce). Protein (150 μg) was hybridized on the antibody arrays overnight at 4 °C at a dilution of 1:5,000. IRDye-labelled streptavidin (LI-COR) was used for the detection, slides were scanned using a Odyssey CLx scanner (LI-COR) and analysed using Image Studio 2.0 software. The results were then normalized using internal controls, and the relative protein levels determined across four biological replicates.

Senescence β-galactosidase staining. A375 cells were grown in low-serum media and treated with vehicle or vemurafenib (0.1 μM) for 3 or 8 days, β-galactosidase staining was performed according to the manufacturer's instructions (Cell Signaling). All experiments were performed independently three times. Representative experiments are shown.

Statistical analysis. Data are generally expressed as mean ± s.e.m., or in box plots in which the centre line is the median, and whiskers are minimum to maximum values. Group sizes were determined based on the results of preliminary experiments and no statistical method was used to predetermine sample size. Group allocation and outcome assessment were not performed in a blinded manner. All samples that met proper experimental conditions were included in the analysis. Statistical significance was determined using a two-tailed Mann–Whitney U test or Student's t-test using Prism 6 software (GraphPad Software), or using a hypergeometric variability test (http://www.geneprof.org). Significance was set at $P < 0.05$.

31. Tavazoie, S. F. *et al.* Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* **451**, 147–152 (2008).
32. Doyle, J. P. *et al.* Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell* **135**, 749–762 (2008).
33. Zhang, X. H. *et al.* Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. *Cell* **154**, 1060–1073 (2013).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

35. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

36. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).

37. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6,** e21800 (2011).

38. Valiente, M. *et al.* Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell* **156,** 1002–1016 (2014).

LETTER

Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer

Christina Twyman-Saint Victor[1,2]*, Andrew J. Rech[2]*, Amit Maity[3,4], Ramesh Rengan[3,4]†, Kristen E. Pauken[5,6], Erietta Stelekati[5,6], Joseph L. Benci[2,3], Bihui Xu[2,3], Hannah Dada[2,3], Pamela M. Odorizzi[5,6], Ramin S. Herati[1,6], Kathleen D. Mansfield[5,6], Dana Patsch[3], Ravi K. Amaravadi[1,4], Lynn M. Schuchter[1,4], Hemant Ishwaran[7], Rosemarie Mick[4,8], Daniel A. Pryma[4,9], Xiaowei Xu[4,10], Michael D. Feldman[4,10], Tara C. Gangadhar[1,4], Stephen M. Hahn[3,4]†, E. John Wherry[4,5,6]§, Robert H. Vonderheide[1,2,4,6]§ & Andy J. Minn[2,3,4,6]§

Immune checkpoint inhibitors[1] result in impressive clinical responses[2–5], but optimal results will require combination with each other[6] and other therapies. This raises fundamental questions about mechanisms of non-redundancy and resistance. Here we report major tumour regressions in a subset of patients with metastatic melanoma treated with an anti-CTLA4 antibody (anti-CTLA4) and radiation, and reproduced this effect in mouse models. Although combined treatment improved responses in irradiated and unirradiated tumours, resistance was common. Unbiased analyses of mice revealed that resistance was due to upregulation of PD-L1 on melanoma cells and associated with T-cell exhaustion. Accordingly, optimal response in melanoma and other cancer types requires radiation, anti-CTLA4 and anti-PD-L1/PD-1. Anti-CTLA4 predominantly inhibits T-regulatory cells (T_{reg} cells), thereby increasing the CD8 T-cell to T_{reg} (CD8/T_{reg}) ratio. Radiation enhances the diversity of the T-cell receptor (TCR) repertoire of intratumoral T cells. Together, anti-CTLA4 promotes expansion of T cells, while radiation shapes the TCR repertoire of the expanded peripheral clones. Addition of PD-L1 blockade reverses T-cell exhaustion to mitigate depression in the CD8/T_{reg} ratio and further encourages oligoclonal T-cell expansion. Similarly to results from mice, patients on our clinical trial with melanoma showing high PD-L1 did not respond to radiation plus anti-CTLA4, demonstrated persistent T-cell exhaustion, and rapidly progressed. Thus, PD-L1 on melanoma cells allows tumours to escape anti-CTLA4-based therapy, and the combination of radiation, anti-CTLA4 and anti-PD-L1 promotes response and immunity through distinct mechanisms.

Anecdotal clinical reports suggest that radiation may cooperate with anti-CTLA4 to systemically enhance melanoma response[7]; however, this combination has not been reported in a clinical trial. To examine the feasibility and efficacy of radiation combined with immune checkpoint blockade, we initiated a phase I clinical trial of 22 patients with multiple melanoma metastases (Extended Data Table 1). A single index lesion was irradiated with hypofractionated radiation, delivered over two or three fractions, followed by four cycles of the anti-CTLA4 antibody ipilimumab (Extended Data Fig. 1a). Accrual was completed in three out of four radiation dose levels, and treatment was well tolerated (Extended Data Table 2). Evaluation of the unirradiated lesions by computed tomography (CT) imaging using response evaluation criteria in solid tumours (RECIST) demonstrated that 18% of patients had a partial response as best response, 18% had stable disease, and 64% had

progressive disease (Fig. 1a). For example, patient PT-402 showed a large reduction in sizes of unirradiated tumours and a partial metabolic response by positron emission tomography (PET) (Fig. 1b). None of the 12 patients evaluated by PET had progressive metabolic disease in the irradiated lesion (Extended Data Fig. 1b, Extended Data Table 3). The median progression-free survival and overall survival was 3.8 and 10.7 months with median follow-up of 18.4 and 21.3 months (18.0 and 21.3 for patients without event), respectively (Fig. 1c).

Although responses were observed, the majority of patients in our trial did not respond. To understand the contribution of radiation to immune checkpoint blockade and to discover mechanisms of resistance, we used the B16-F10 melanoma mouse model. Mice with bilateral flank tumours received anti-CTLA4, irradiation of one tumour (index) using a micro-irradiator, or both treatments delivered concurrently (Fig. 1d). The best responses in both tumours occurred with radiation + anti-CTLA4. Radiation given before or concurrently with CTLA4 blockade yielded similar results (Extended Data Fig. 1c). Complete responses were CD8 T-cell-dependent, and mice with complete responses also exhibited CD8 T-cell-dependent immunity to tumour re-challenge (Extended Data Fig. 1d–e). However, similar to our clinical trial, only approximately 17% of mice responded. To better understand determinants of response, we derived cell lines from unirradiated tumours that relapsed after radiation + anti-CTLA4 (Res 499 and Res 177). Resistance was confirmed *in vivo* and was not due to intrinsic radiation resistance (Extended Data Fig. 2a–c). Random forest machine learning analysis[8,9] of tumour-infiltrating lymphocytes (TILs) demonstrated that the top predictor of resistance, as measured by variable importance scores and selection, was the $CD8^+CD44^+$ to T_{reg} (CD8/T_{reg}) ratio (Fig. 1e, Extended Data Fig. 2d). In resistant tumours, the CD8/T_{reg} ratio failed to increase after radiation + anti-CTLA4 as it did in sensitive tumours because $CD8^+CD44^+$ T cells did not significantly expand despite reduction in T_{reg} cells (Fig. 1f). Other immune variables associated with resistance were also related to the failure to accumulate CD8 TILs.

The prevalence of CD8 TILs can be blunted by mechanisms that interfere with T-cell function. Transcriptomic profiling of Res 499/177 tumours revealed that PD-L1 was among the top 0.2% of upregulated genes that make up a radiation + anti-CTLA4 'resistance gene signature' (Extended Data Fig. 2e, Supplementary Table 1). Other genes include interferon-stimulated genes, which may promote immune suppression through PD-L1[10,11]. Similarly, PD-L1 was co-expressed with the

[1]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [2]Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [3]Department of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [4]Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [5]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [6]Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [7]Division of Biostatistics, Department of Public Health Sciences, University of Miami, Miami, Florida 33136, USA. [8]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [9]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [10]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. †Present addresses: Department of Radiation Oncology, University of Washington School of Medicine, Washington 98195, USA (R.R.); Division of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA (S.M.H.).
*These authors contributed equally to this work.
§These authors jointly supervised this work.
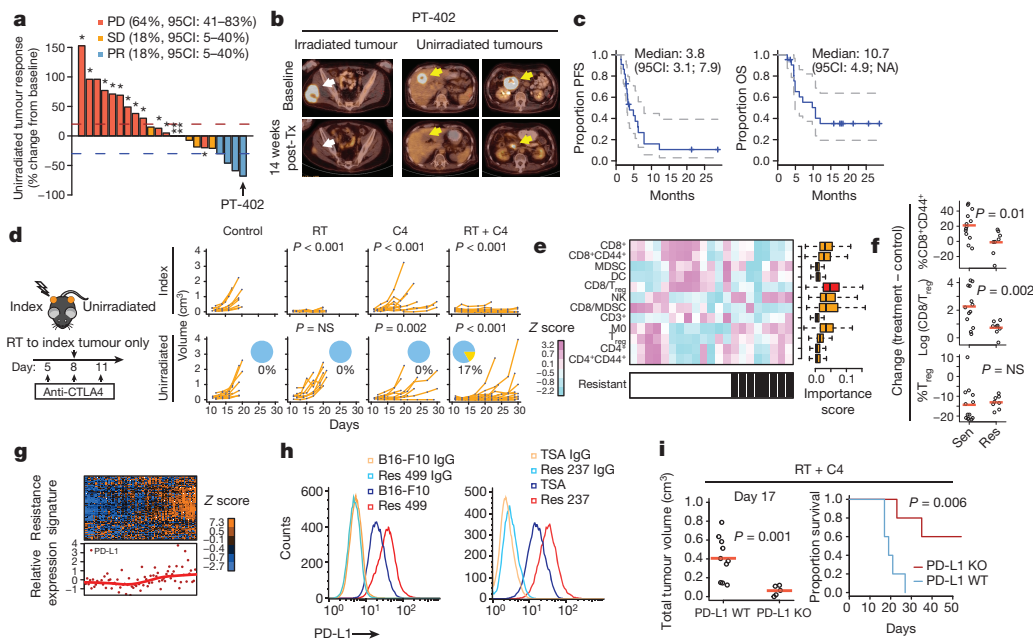
Figure 1 | Radiation + anti-CTLA4 promotes regression of irradiated and unirradiated tumours and is inhibited by PD-L1 on tumour cells.
a, Waterfall plot of unirradiated tumours after radiation treatment (RT) to a single index lesion with anti-CTLA4. Dashed lines are thresholds for progressive disease (PD; red) and partial response (PR; blue). *Patients with new lesions. **Clinical progression without imaging. 95CI, 95% confidence interval. **b**, PET/CT images of irradiated (white arrows) and unirradiated (yellow arrows) tumours from patient PT-402. **c**, Progression-free survival (PFS) and overall survival (OS) for all patients (dashed lines, 95CI). **d**, B16-F10 tumour growth after RT to the index tumour ($n = 8$), anti-CTLA4 (C4) ($n = 9$), anti-CTLA4 and RT to the index tumour ($n = 18$), or no (control) treatment ($n = 9$). The P values are comparisons with control using a linear mixed-effects model. Pie chart shows per cent complete responses (yellow).

See Fig. 2d for survival. **e**, Heat map showing relative abundance of immune cells or their ratios from tumours that are resistant (black hatch) or sensitive to RT + anti-CTLA4. Boxplot shows bootstrap importance scores for each variable. Higher values (red) are more predictive. **f**, Change in T cell subsets or their ratio after RT + anti-CTLA4 for sensitive parental (Sen) or resistant (Res) tumours. Values are subtracted from average of untreated controls. Red line is mean. **g**, Heat map of resistance gene signature and PD-L1 across human melanoma. $P < 0.001$ by gene set enrichment analysis. **h**, Expression of PD-L1 on Res 499 compared to B16-F10 melanoma cells and of Res 237 compared to TSA breast cancer cells. Isotype control (IgG). **i**, Total tumour volume from PD-L1 knockout (KO) or control (WT) Res 499 and corresponding survival. Two-tailed t-test or Wilcoxon test was used for two-way comparisons of biological replicates. Log-rank test was used for survival analysis.
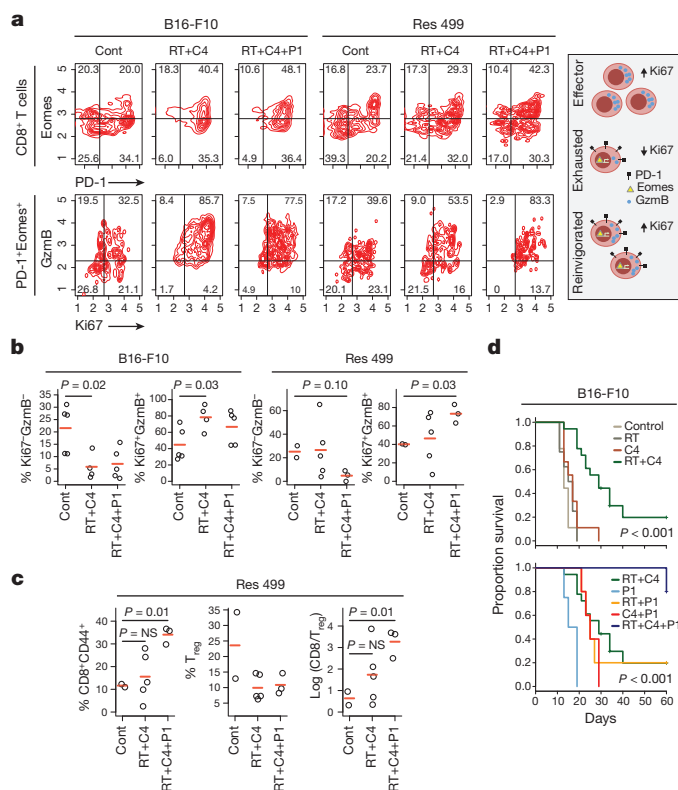
resistance signature in tumours from a previously reported[12] cohort of metastatic melanoma patients (Fig. 1g). This increase in PD-L1 was observed on melanoma cells devoid of contaminating stromal cells, and a comparable increase was similarly seen in the Res 237 murine breast cancer cell line (Fig. 1h), which was selected from the TSA line for resistance to radiation + anti-CTLA4 (Extended Data Fig. 2f, g). In contrast, expression of other inhibitory receptors and their ligands nominated by gene profiling did not suggest an obvious role in resistance (Extended Data Fig. 2h, i). Indeed, genetic elimination of PD-L1 on Res 499 cells by CRISPR (Extended Data Fig. 2j) restored response to radiation + anti-CTLA4 by increasing survival from 0% to 60% (Fig. 1i). Thus, an increase in PD-L1 on tumour cells observed in multiple cancer types can be a dominant resistance mechanism to radiation + anti-CTLA4.

Elevated levels of PD-L1 can promote T-cell exhaustion, a state characterized by dysfunction in T-cell proliferation and effector function[13]. Exhausted T cells co-express the PD-L1 receptor PD-1 and the

Figure 2 | Addition of PD-L1 blockade reinvigorates exhausted T cells and improves response to radiation + anti-CTLA4. **a**, Representative contour plot of CD8 TILs from B16-F10 or Res 499 tumours after radiation treatment (RT) and anti-CTLA4 (C4) ± anti-PD-L1 (P1) examined for PD-1 and Eomes (top row), followed by examination of the PD-1$^+$Eomes$^+$ subset for Ki67 and GzmB (bottom row). Schema shows exhaustion and reinvigoration markers. **b**, Proportion of PD-1$^+$Eomes$^+$ CD8 T cells that are either Ki67$^-$GzmB$^-$ or Ki67$^+$GzmB$^+$. **c**, Changes in T cell subsets and their ratio from Res 499 tumours. **d**, Survival of mice with B16-F10 tumours ($n = 18$ for RT + C4, $n = 5$ for others). Shown are overall log-rank P values. Two-tailed t-test or Wilcoxon test was used for two-way comparisons of biological replicates.

transcription factor Eomes[14]. Reversal of exhaustion, known as reinvigoration, is marked by an increase in the proliferation marker Ki67 and the cytotoxic protein GzmB within the exhausted T-cell pool. In both untreated parental and resistant tumours, approximately 20% of CD8 TILs co-expressed PD-1 and Eomes, and only a minority of these cells were Ki67[+]GzmB[+], indicating that a significant fraction was exhausted (Fig. 2a, b). In B16-F10 tumours, radiation + anti-CTLA4 markedly increased both the proportion of PD-1[+]Eomes[+] CD8 T cells and the proportion that were Ki67[+]GzmB[+] within this subset. In contrast, in resistant tumours the average proportion of PD-1[+]Eomes[+] T cells that were Ki67[+]GzmB[+] only marginally increased after radiation + anti-CTLA4; however, addition of anti-PD-L1 increased this to levels observed in parental tumours treated with only radiation + anti-CTLA4. The frequency of CD8[+]CD44[+] TILs and the CD8/T_{reg} ratio also increased (Fig. 2c), and these were strongly correlated with the proportion of PD-1[+]Eomes[+] CD8 TILs that were Ki67[+]GzmB[+] (Extended Data Fig. 3a). Importantly, addition of anti-PD-L1 improved responses of resistant Res 499 tumours after radiation + anti-CTLA4 (Extended Data Fig. 3b, c). For treatment-naive tumours, responses were even more notable as the addition of either anti-PD-L1 or anti-PD-1 to radiation + anti-CTLA4 markedly improved survival and increased complete responses to 80% (Fig. 2d, Extended Data Fig. 3d–f). On average, 58% of mice with complete responses after adding anti-PD-L1 or anti-PD-1 were alive 90+ days after tumour rechallenge, and similar improvements were observed with Res 237 breast cancer tumours after addition of PD-L1 blockade (Extended Data Fig. 3g–i). Thus, elevated PD-L1 on tumour cells results in persistent T-cell exhaustion that impairs the CD8/T_{reg} ratio. Addition of PD-L1 blockade inhibits resistance and results in long-term immunity.

Notably, radiation is needed to achieve high complete response rates as dual checkpoint blockade proved inferior to dual checkpoint blockade plus radiation (Fig. 2d), a requirement additionally seen in a pancreatic cancer model (Extended Data Fig. 3j). The superiority of triple therapy in multiple cancer types suggests non-redundant mechanisms for each treatment. To examine this notion, we assessed treatment-related changes in TILs from unirradiated tumours. Random forest modelling of immune cell profiles confirmed that anti-CTLA4 predominantly caused a decrease in T_{reg} cells, anti-PD-L1 strongly increased CD8 TIL frequency, and the blockade of both increased the CD8/T_{reg} ratio (Fig. 3a, b, Extended Data Fig. 4a). In contrast, radiation caused only a modest increase in CD8 TILs; however, TCR sequencing revealed that this was accompanied by increased diversity of TCR clonotypes, which could be observed even in the presence of CTLA4 blockade (Fig. 3c, d). Thus, within the tumour microenvironment, CTLA4 blockade primarily decreases T_{reg} cells, PD-L1 blockade predominantly reinvigorates exhausted CD8 TILs, and radiation diversifies the TCR repertoire of TILs from unirradiated tumours.

To investigate if treatment effects on TILs were propagated to the peripheral T-cell pool, we examined spleen and blood. As observed in TILs, radiation + anti-CTLA4 reinvigorated exhausted PD-1[+]Eomes[+] splenic CD8 T cells, and this reinvigoration was further enhanced by
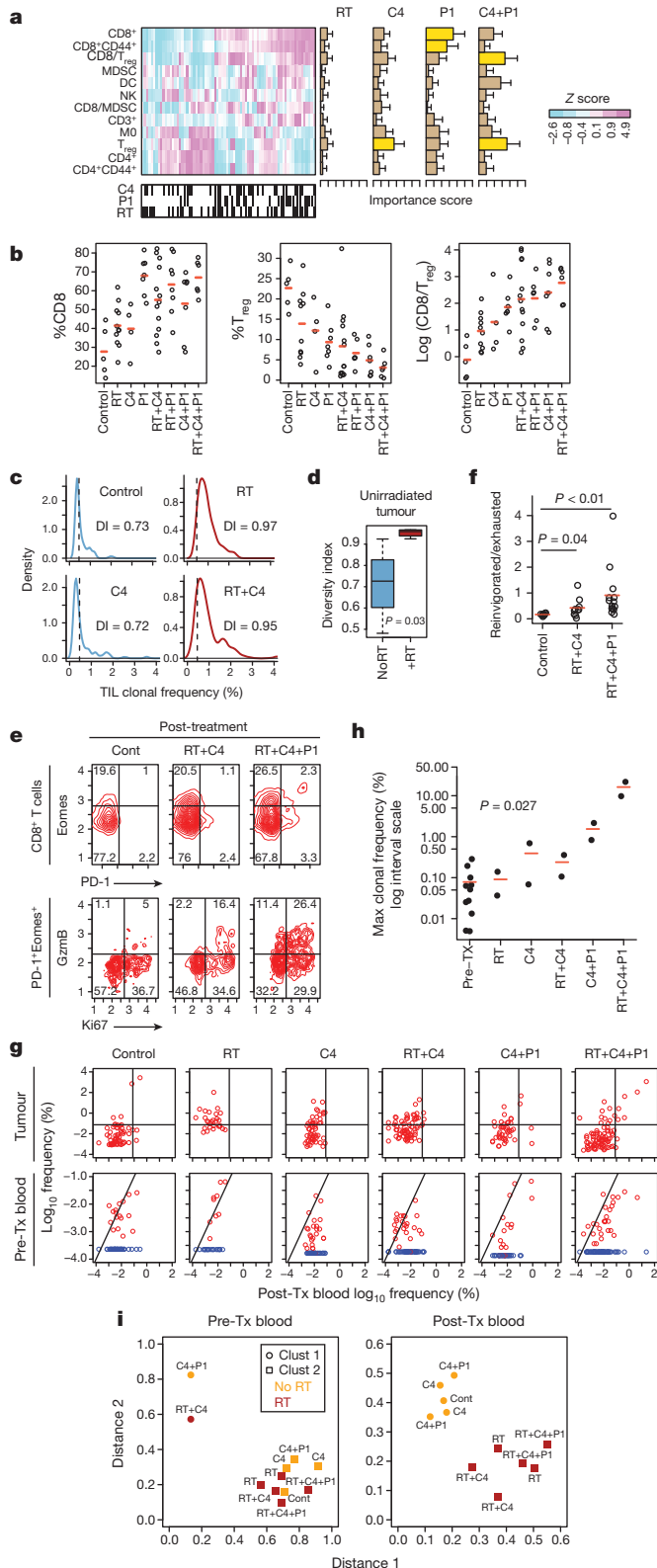

Figure 3 | Radiation, anti-CTLA4, and anti-PD-L1 have distinct effects on the TCR repertoire, T_{reg} cells, and T-cell exhaustion. a, Heat map of changes in the frequency of immune cells or their ratios from B16-F10 tumours. Black hatches indicate treatment. Bar plots show bootstrap importance scores (mean ± s.e.m.) that assess changes in immune parameters predicted by treatment type (read row-wise). Higher values (yellow) represent stronger association. RT, radiation treatment; C4, anti-CTLA4; P1, anti-PD-L1. **b**, T cell subsets and their ratios. **c**, Frequency distribution (dashed line is 0.5%) and **d**, boxplot of diversity index (DI; 0, clonal; 1, fully diverse) for most frequent TCR clonotypes found in TILs of unirradiated B16-F10 tumours after RT and/or anti-CTLA4. Boxplot summarizes data for mice treated with anti-CTLA4 (NoRT) or RT ± anti-CTLA4 (+RT). **e**, Representative contour plots and **f**, ratios examining PD-1[+]Eomes[+] splenic CD8 T cells from mice with B16-F10 tumours for Ki67[+]GzmB[+] (reinvigorated) or Ki67[−]GzmB[−] (exhausted) subsets. **g**, TCR clonal frequency in post-treatment blood vs TILs (top row) or vs pre-treatment blood (bottom row). Quadrant boundaries are top 5% quantiles from the control. Clones below detection in pre-treatment blood are assigned upper bounds (blue). **h**, Maximum clonal frequency in post-treatment blood (dot) of the most frequent TCR clonotypes found in TILs. P value by Kruskal-Wallis test. **i**, Distances to cluster centroids for the average CDR3 amino acid features of the five most frequent clones in pre- or post-treatment blood from mice treated with (red) or without (orange) RT. Membership into two clusters (circles and squares) determined by k-means. Two-tailed t-test or Wilcoxon test was used for two-way comparisons.

addition of anti-PD-L1 (Fig. 3e, f). Reinvigoration after addition of anti-PD-L1 was also accompanied by a large expansion of a small subset of the top 100 most frequent TCR clonotypes found in TILs (Fig. 3g). Remarkably, some clones reached a frequency in the post-treatment blood of over 20% after radiation and dual checkpoint blockade (Fig. 3h). With anti-CTLA4 ± radiation, peripheral T cell clonal expansion was modest, which parallels the low complete response rates following this treatment. Radiation alone was insufficient to drive peripheral T-cell expansion, despite increasing TCR repertoire diversity of TILs, but did promote qualitative alterations in the TCR repertoire of the most expanded clonotypes. Unsupervised analysis using the average CDR3 amino acid features[15,16] demonstrated that the TCRs of the most frequent clonotypes in the post-treatment blood formed two readily apparent clusters on the basis of radiation treatment (Fig. 3i). In contrast, the most frequent clonotypes from pre-treatment blood and randomly sampled clonotypes from post-treatment blood did not separate into clusters, consistent with differences in CDR3 amino acid properties being an effect of radiation only observed in the most expanded clones (Extended Data Fig. 4b, c). The separation into two clusters was driven by differences in the CDR3 occupancy profile of short amino acid sequences belonging to distinct subsets differing in size, polarity, and electrostatic charge (Extended Data Fig. 4d, e). Together, these observations suggest that the favourable immune changes in TILs after immune checkpoint blockade promote their peripheral clonal expansion. When combined with increased TCR repertoire diversity afforded by radiation, selection and oligoclonal peripheral expansion of clones with distinct TCR traits are favoured.

To determine if treatment and resistance-related changes in peripheral T cells can constitute a biomarker for tumour response, we modelled the effects of reinvigoration, exhaustion, and the CD8/T_{reg} ratio. Specifically, we used (1) the percentage of PD-1$^+$ splenic CD8 T cells that are Eomes$^+$ to integrate the burden that exhausted T cells might exert, (2) the percentage of PD-1$^+$ CD8 T cells that are Ki67$^+$GzmB$^+$ as a measure of potential reinvigoration, and (3) the CD8/T_{reg} ratio as a barometer for the suppressive potential of T_{reg} cells. The overall prediction accuracy of the model was 84%, and variables for T-cell reinvigoration and exhaustion were the most predictive, followed by the CD8/T_{reg} ratio (Extended Data Fig. 5a, b). Moreover, the percentage of PD-1$^+$ CD8 T cells that were Eomes$^+$ was a striking modifier of the likelihood of complete response as nearly all observed complete responses occurred when the percentage of Ki67$^+$GzmB$^+$ in PD-1$^+$ CD8 T cells was high but the relative size of the PD-1$^+$Eomes$^+$ exhausted population was small (Fig. 4a). Similar relationships existed with the CD8/T_{reg} ratio, and prediction using T cells from peripheral blood yielded highly similar results (Extended Data Fig. 5c–e). In total, immune parameters from peripheral T cells that relate the size of the exhausted T-cell population, reinvigoration, and the CD8/T_{reg} ratio can predict response to radiation combined with immune checkpoint blockade.

To assess whether immune predictors discovered in mice could be shared with patients, we examined peripheral T cells and tumour biopsies from patients on our clinical trial of radiation + anti-CTLA4. For all 10 patients with available pre- and post-treatment blood, two had partial responses in unirradiated tumours and progression-free survival significantly longer than the median. For both of these patients, the percentages of Ki67$^+$GzmB$^+$ increased in PD-1$^+$Eomes$^+$ CD8 T cells after treatment while the proportion of PD-1$^+$Eomes$^+$ T cells remained at or below the mean (Fig. 4b). In contrast, patients with a high percentage of PD-1$^+$Eomes$^+$ T cells post-treatment did not have partial responses and had a short progression-free survival, regardless of reinvigoration. Comparison of patient PT-402, who had extended progression-free survival/partial response (Fig. 1a, b), with patient PT-102, who had short progression-free survival/progressive disease, demonstrates how reinvigoration is associated with response to radiation + anti-CTLA4 as it is in mice (Fig. 4c vs Fig. 3e, f and Extended Data Fig. 5f, g). Examination of pre-treatment tumour biopsies from patients PT-402 and PT-102 (Fig. 4d), and from all patients with available biopsy (Extended Data
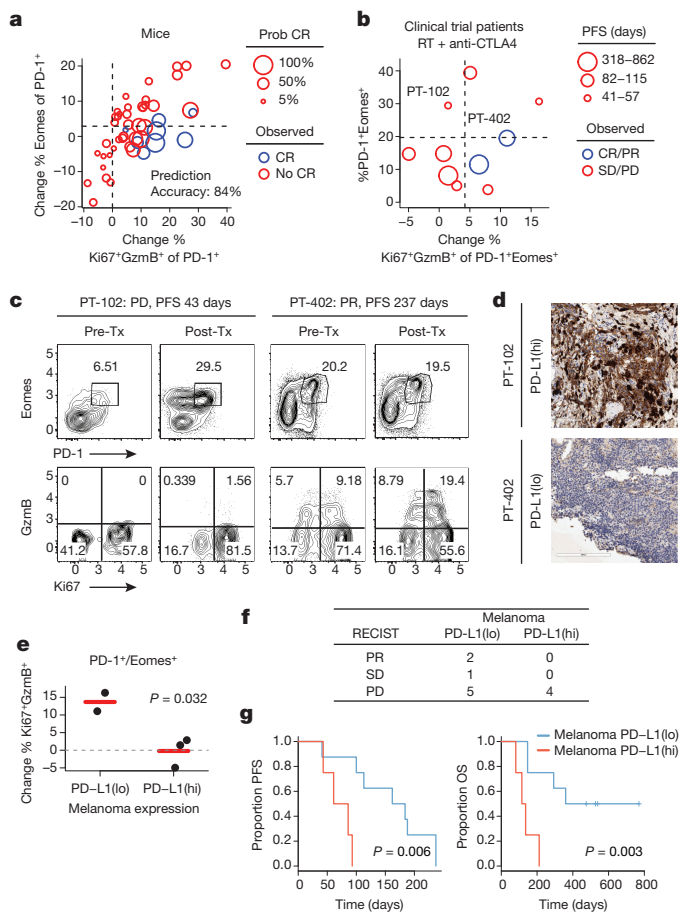
Figure 4 | Tumour PD-L1 and T-cell exhaustion and reinvigoration can predict response in mice and patients. a, Percentage of PD-1$^+$ CD8 T cells that are Eomes$^+$ vs Ki67$^+$GzmB$^+$ after radiation treatment (RT) combined with checkpoint blockade. Values are subtracted from average of untreated controls. Each circle represents a mouse. Probability of complete response (CR; proportional to circle size), prediction error rate, and quadrant boundaries are estimated from a random forest model. b, Percentage of Eomes$^+$PD-1$^+$ CD8 T cells in post-treatment blood vs change in % PD-1$^+$Eomes$^+$ CD8 T cells that are Ki67$^+$GzmB$^+$ after treatment. Each circle represents a patient. Progression-free survival (PFS) is proportional to circle size and quadrant boundaries are average values for patients under the mean PFS. Concordance index of the random forest model is 0.59. c, Contour plot of peripheral blood CD8 T cells from patients PT-102 and PT-402 examined for PD-1 and Eomes (top row), followed by examination of the PD-1$^+$Eomes$^+$ subset for Ki67 and GzmB (bottom row). d, PD-L1 staining from corresponding tumour biopsies. e, Change in per cent Ki67$^+$GzmB$^+$ in PD-1$^+$Eomes$^+$ CD8 T cells vs PD-L1 status of melanoma cells from all patients with available pre- and post-treatment blood. f, RECIST response; g, PFS and overall survival (OS) stratified by PD-L1 status of melanoma cells.

Table 4), revealed that PD-L1^{lo} intensity on melanoma cells (Extended Data Fig. 6a) was associated with reinvigoration of PD-1$^+$Eomes$^+$ and of PD-1$^+$ CD8 T cells after radiation + anti-CTLA4, while PD-L1^{hi} status was associated with persistent exhaustion (Fig. 4e, Extended Data Fig. 6b). None of the patients with PD-L1^{hi} on melanoma cells had a complete response/partial response, and all rapidly progressed and died (Fig. 4f, g). PD-L1 status on macrophages was neither associated with reinvigoration nor independently predictive of progression-free survival (Extended Data Fig. 6c, d). Thus, collective results from patients and mice suggest that elevated PD-L1 on melanoma cells inhibits T-cell function and tumour response to radiation + anti-CTLA4.

We investigated radiation + anti-CTLA4 in mice and patients to understand mechanisms of both response and resistance (Extended Data Fig. 6e). Anti-CTLA4 predominantly inhibits T_{reg} cells, increasing the

$CD8/T_{reg}$ ratio as previously described[17], and results in modest peripheral expansion of TCR clonotypes in the tumour, also consistent with other reports[18,19]. Radiation diversifies the TCR repertoire of TILs and shapes the repertoire of expanded clones. Although the cause and consequence of these repertoire changes remain to be defined, radiation can alter peptide presentation[20], and CDR3 changes after *Mycobacterium tuberculosis* infection have been hypothesized to be antigen-driven[15]. Resistance to radiation + anti-CTLA4 can ensue due to elevated PD-L1 on cancer cells driving T-cell exhaustion, a process that can be antagonized by PD-L1 blockade. However, severely exhausted T cells may regain only limited function after reinvigoration[13,14], explaining why the correlation between reinvigoration and response declines when the exhausted T-cell pool is large. Although tumours with genetic elimination of PD-L1 in melanoma cells can still relapse, suggesting resistance through other pathways and/or PD-L1 on non-tumour cells, the upregulation of PD-L1 by cancer cells is a dominant resistance mechanism in our models. Moreover, the shared findings between mice and patients predict that addition of PD-L1/PD-1 blockade to radiation + anti-CTLA4 may show significant efficacy in clinical trials.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nature Rev. Cancer* **12,** 252–264 (2012).
2. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
3. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369,** 134–144 (2013).
4. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
5. Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366,** 2455–2465 (2012).
6. Wolchok, J. D. *et al.* Nivolumab plus ipilimumab in advanced melanoma. *N. Engl. J. Med.* **369,** 122–133 (2013).
7. Postow, M. A. *et al.* Immunologic correlates of the abscopal effect in a patient with melanoma. *N. Engl. J. Med.* **366,** 925–931 (2012).
8. Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. High-dimensional variable selection for survival data. *J. Am. Stat. Assoc.* **105,** 205–217 (2010).
9. Ishwaran, H., Kogalur, U. B., Chen, X. & Minn, A. J. Random survival forests for high-dimensional data. *Statistical Analy Data Mining* **4,** 115–132 (2011).
10. Wilson, E. B. *et al.* Blockade of chronic type I interferon signaling to control persistent LCMV infection. *Science* **340,** 202–207 (2013).
11. Teijaro, J. R. *et al.* Persistent LCMV infection is controlled by blockade of type I interferon signaling. *Science* **340,** 207–211 (2013).
12. Jönsson, G. *et al.* Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin. Cancer Res.* **16,** 3356–3367 (2010).
13. Wherry, E. J. T cell exhaustion. *Nature Immunol.* **12,** 492–499 (2011).
14. Paley, M. A. *et al.* Progenitor and terminal subsets of CD8$^+$ T cells cooperate to contain chronic viral infection. *Science* **338,** 1220–1225 (2012).
15. Thomas, N. *et al.* Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30,** 3181–3188 (2014).
16. Atchley, W. R., Zhao, J., Fernandes, A. D. & Druke, T. Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA* **102,** 6395–6400 (2005).
17. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. PD-1 and CTLA-4 combination blockade expands infiltrating T cells and reduces regulatory T and myeloid cells within B16 melanoma tumors. *Proc. Natl Acad. Sci. USA* **107,** 4275–4280 (2010).
18. Cha, E. *et al.* Improved survival with T cell clonotype stability after anti-CTLA-4 treatment in cancer patients. *Sci. Transl. Med.* **6,** 238ra70–238ra70 (2014).
19. Robert, L. *et al.* CTLA4 blockade broadens the peripheral T-cell receptor repertoire. *Clin. Cancer Res.* **20,** 2424–2432 (2014).
20. Reits, E. A. Radiation modulates the peptide repertoire, enhances MHC class I expression, and induces successful antitumor immunotherapy. *J. Exp. Med.* **203,** 1259–1271 (2006).

Author Contributions C.T. and A.J.R. designed, performed, and/or analysed mouse and *in vitro* experiments. H.D. assisted with mouse experiments. J.L.B. performed genetic ablation experiments for PD-L1. B.X. assisted with breast cancer mouse model. K.E.P. and P.M.O. designed and/or performed mouse immune profiling studies. E.S., R.S.H. and K.D.M. designed and/or performed human immune profiling studies. R.R. designed the clinical trial and was principle investigator along with S.M.H. and A.M., who oversaw its completion. D.P. assisted in coordinating the trial. A.M. evaluated the final clinical trial data. R.K.A., T.C.G. and L.M.S. were investigators on the trial, D.A.P. evaluated all imaging response for trial patients and assisted in study design, and M.D.F. and X.X. evaluated pathological biomarkers. R.M. was responsible for the statistical design of the clinical trial, and R.M. and A.J.M. performed statistical analysis. H.I. and A.J.M. performed and interpreted statistical analysis of pre-clinical data. A.J.M. wrote the manuscript, and E.J.W., R.H.V., A.M., C.T. and A.J.R. edited the manuscript. A.J.M., E.J.W. and R.H.V. together designed, interpreted, and oversaw the study.

Author Information The transcriptomic data are available at the GEO repository (http://www.ncbi.nlm.nih.gov/geo/) under accession GSE65503. Code for computational analysis is available upon request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.J.M. (andyminn@exchange.upenn.edu).

METHODS

Clinical trial patients and study design. The clinical protocol was registered on http://clinicaltrials.gov (NCT01497808). Eligible patients were at least 18 years of age with previously treated or untreated stage IV melanoma with multiple metastasis. Patients were required to have an Eastern Cooperative Oncology Group performance status of 0 or 1, adequate renal, hepatic, and haematological function, no current or history of CNS metastasis, no prior radiation that precludes use of stereotactic body radiation (SBRT), and at least one tumour between 1 and 5 cm that could be treated with SBRT. The primary objectives of this phase I study were to determine feasibility, dose-limiting toxicities (DLT) and maximum tolerated SBRT fraction when given in conjunction with ipilimumab. The secondary objectives were to determine late toxicity, immune-related clinical responses and changes. The study treated successive cohorts of patients with escalating doses of SBRT to a single tumour (index lesion), followed 3–5 days later by ipilimumab every three weeks for four doses. Moderate radiation doses were used since higher radiation dose has not been clearly correlated with better immune response but would be likely to increase toxicity. Patients were stratified into two strata based on treatment site (lung or bone vs liver or subcutaneous) and dose escalation of SBRT was determined as follows: For lung/bone lesion, dose level 1 (DL1) was 8 Gy × 2; dose level 2 (DL2) was 8 Gy × 3; and for liver/subcutaneous lesion, DL1 was 6 Gy × 2; DL2 was 6 Gy × 3. The study followed a "treat six" design with the goal of accruing 6 patients to each dose level, or 24 patients total. Enrolment to a dose level would stop if 2 or more patients had a DLT. If 0-1 patients out of the 6 had a DLT at DL1, escalation to DL2 would proceed. There were no observed DLTs, defined by the protocol as any treatment-related grade 4 or higher immune-related toxicity (NCI CTC Version 4.0) or grade 3 or higher non-immune related toxicity experienced during study treatment or within 30 days after the last injection of ipilimumab. Pre- and post-treatment blood, CT, and PET/CT scans were obtained to follow tumour response and assess immune responses. Response evaluation by imaging was performed within 60 days of the last ipilimumab treatment using either RECIST v1.1[21] or PERCIST. The study protocol was approved by the University of Pennsylvania institutional review board. All participating patients provided written informed consent.

Cell lines and tissue culture. B16-F10 was purchased from ATCC. TSA was a gift from Sandra Demaria. PDA.4662 cell line was derived from single-cell suspensions of PDA tissue from Kras$^{LSL-G12D/+}$, p53$^{LSL-R172H/+}$,Pdx1-Cre mice as previously described[22]. B16-F10 and PDA.4662 cell lines were cultured at 37 °C in DMEM and TSA cells were cultured at 37 °C in RPMI. Media was supplemented with 10% FBS, 100 U ml^{-1} penicillin and 100 µg ml^{-1} streptomycin, 2 mM L-glutamine. All cell lines were determined to be free of Mycoplasma (Lonza) and common mouse pathogens (IDEXX).

***In vivo* mouse studies.** Five to seven week old female C57BL/6 and BALB/c mice were obtained from NCI Production (Frederick, MD) and Jackson Laboratory (Bar Harbour, ME) and maintained under pathogen-free conditions. All animal experiments were performed according to protocols approved by the Institute of Animal Care and Use Committee of the University of Pennsylvania. For B16-F10 melanoma, 5×10^4 B16-F10 cells were mixed with an equal volume of Matrigel (BD Biosciences) and subcutaneously injected on the right flank of C57BL/6 mice on day 0 and the left flank on day 2. The right flank tumour site was irradiated with 20 Gy on day 8. Blocking antibodies were given on days 5, 8 and 11. For the concurrent vs sequential radiation experiment, the right flank was irradiated on either day 8 (sequential) or 12 (concurrent), while blocking antibodies were given on days 9, 12, and 15. For TSA breast cancer, 1×10^5 TSA cells were mixed with an equal volume of Matrigel (BD Biosciences) and subcutaneously injected on the right flank of BALB/c mice on day 0 and the left flank on day 2. The right flank was irradiated with 8 Gy on three consecutive days starting on day 10 or 11 post tumour implantation. Blocking antibodies were started 3 days before radiation and given every 3 days for a total of 3 doses. For the pancreatic cancer model, 4×10^5 PDA.4662 cells were subcutaneously injected on the right flank. The right flank was irradiated with 20 Gy on day 8. Blocking antibodies were given on days 5, 8, and 11. For melanoma and breast cancer models, we used the optimal dose and fraction of radiation as previously reported[23,24]. All irradiation was performed using the Small Animal Radiation Research Platform (SARRP). Antibodies used for *in vivo* immune checkpoint blockade experiments were given intraperitoneally at a dose of 200 µg per mouse and include: CTLA4 (9H10), PD-1 (RMP1-14), PD-L1 (10F.9G2), CD8 (2.43), and rat IgG2B isotype (LTF-2) (BioXCell). Anti-CD8 was given 2 days before tumour implantations (day −2), day 0, then every 4 days for the duration of the experiment. Perpendicular tumour diameters were measured using calipers. Volume was calculated using the formula $L \times W^2 \times 0.52$, where L is the longest dimension and W is the perpendicular dimension.

Survival and tumour response analysis. Differences in survival were determined for each group by the Kaplan–Meier method and the overall P value was calculated by the log-rank test using the "survival" R package version 2.37+. For mouse studies, an event was defined as death or when tumour burden reached a protocol-specified size of 1.5 cm in maximum dimension to minimize morbidity. To help control for differences in treatment response due to experimental variation or intrinsic growth differences with sublines, tumour volume measurements were also analysed after normalizing to the average volumes of untreated control mice. These average untreated tumour volumes were determined at day 11–12, a time when tumour dimensions could be accurately measured, and was considered a baseline tumour volume (V_{cont}). Normalized tumour response to treatment is the measured volume (V) relative to V_{cont}, or V/V_{cont}, a dimensionless value. Measurements from different experiments separated by 1–2 days were binned. Differences in growth curves were determined by a linear mixed-effects model with normalized data using the "lmerTest" R package version 2.0. Sample size estimations were based on preliminary pilot experiments. For control mice, we expected an average tumour volume of 0.4 cm^3 at day 17–21. For most experiments, we assumed the treatment group would have an effect size resulting in a 50% reduction in average tumour volume. Sigma was estimated to be 1.5. For a 0.80 power at the 0.05 alpha level, this gave us a sample size of 5 mice. Mice were randomly assigned a treatment group. For experiments whereby the effect size was expected to be small and/or non-robust, two independent researchers with at least one researcher blinded to the treatment group assignments performed caliper measurements.

Flow cytometry. For flow cytometric analysis of *in vivo* experiments, blood, spleen, and tumour were harvested at either day 16 or 18 post tumour implantation. Single-cell suspensions were prepared and red blood cells were lysed using ACK Lysis Buffer (Life Technologies). Live/dead cell discrimination was performed using Live/Dead Fixable Aqua Dead Cell Stain Kit (Life Technologies) or Sytox Red Dead Cell Stain (Life Technologies). Cell surface staining was done for 20–30 min. Intracellular staining was done using a fixation/permeabilization kit (eBioscience.) T effector cells were phenotyped as CD8$^+$CD44$^+$, myeloid derived suppressor cells (MDSC) as CD11b$^+$Gr-1$^+$, and regulatory T cells (T$_{reg}$ cells) as CD4$^+$FOXP3$^+$. All flow cytometric analysis was done using an LSR II (BD) or FACSCalibur (BD) and analysed using FlowJo software (TreeStar) or the FlowCore package in the R language and environment for statistical computing. See Supplementary Methods for a list of antibodies used.

CRISPR gene targeting. Gene targeting by CRISPR/Cas9 was accomplished by co-transfection of a Cas9 plasmid (Addgene, 56503), the guide sequence (selected using ZiFit Targeter) cloned into the gBlock plasmid, and a plasmid with the puromycin selection marker. Successful targeting of PD-L1 was determined by flow cytometry screening of clones treated with and without 100 ng ml^{-1} of interferon (IFN)-gamma (PeproTech). Confirmed clones were pooled. Clones without knockout were also pooled and used as controls. See Supplementary Methods for guide RNA sequences.

Immunohistochemistry for PD-L1. Formalin-fixed, paraffin-embedded tumours were collected at the time of surgical resection or from biopsy. All patients with available recent biopsy, which was optional for trial enrolment, were used for analysis. After heat-induced antigen retrieval (Bond ER2, 20 min.), the tumour slides were stained with an anti-PD-L1 antibody (E1L3N, Cell Signaling) at 1:50 dilution. Intensity of staining on a 0–3+ scale, the percentage of tumour cells or macrophages with positive staining, and the cellular pattern (membrane vs cytoplasm) were analysed by two pathologists. Samples with membrane PD-L1 staining intensity score of 0–1 were classified as PD-L1^{lo}, and samples with an intensity score of 2+ in at least 1% of the cells were classified as PD-L1^{hi}. To confirm specificity, the anti-PD-L1 antibody was validated by staining Hodgkin's lymphoma cells[25] and placenta[26].

Statistical methods and software. Computational analysis and predictions were performed using the R language and environment for statistical computing (version 3.0+) and Bioconductor (version 2.22+). The significance of all two-way comparisons was determined by two-sample, two-tailed t-test. An F-test was used to test for equal variance and a Shapiro–Wilks test was used to test for normality. For non-parametric data, a Wilcoxon test was used. A linear mixed-effects model was used to determine significance of differences in tumour growth. Simple correlation between variables was done using a Pearson's correlation. Unless noted, samples were independent biological replicates.

Microarray data processing and normalization. Total RNA was isolated and purified from cells using Isol-RNA Lysis Reagent (Fisher). Total RNA from tumours was isolated and purified from frozen specimens using Isol-RNA Lysis Reagent and Qiagen RNeasy extraction kit with DNase I on-column treatment. Labelled RNA was hybridized to the Affymetrix GeneChip Mouse Gene 1.0 and 2.0 ST Array. Affymetrix CEL files for all samples were processed using the RMA method as implemented in the "oligo" R package version 1.26.6. Probe annotations were provided by the "mogene10sttranscriptcluster.db" and "mogene20sttranscriptcluster.db" R package version 8.0.1 and 2.13.0, respectively. Since different array types and different batches were used, each expression set was Z-score transformed[27] and median centred. Multiple probes for the same gene were averaged and only genes

common to the 1.0 and 2.0 ST arrays were kept. Batch effects were adjusted using the ComBat method as implemented in the "sva" R package version 3.8.0. The microarray data has been deposited at the GEO (GSE65503) and processed data provided as Supplementary Table 2. Gene expression data for primary melanoma samples were downloaded from the GEO (GSE22155). For this data set, the post-processed data and provided annotations were used.

Determining differentially expressed genes and enriched gene sets. Non-specific filtering was used to remove genes with an interquartile range less than 0.05. To find differentially expressed genes between parental sensitive and resistant tumours, Significance Analysis of Microarray[28] ("samr" R package version 2.0) was applied using a two class unpaired comparison, minimal Z-score fold change of 1.2, and median false discovery rate of 0.05. Unannotated transcripts were not considered. To test whether gene sets were enriched in response to different conditions, we used Gene Set Analysis as implemented in the "GSA" R package version 1.03[29]. The "maxmean" test statistic was used to test enrichment using a two-class comparison. All P values and false discovery rates were based on 500–1,000 permutations. For restandardization, a method that combines randomization and permutation to correct permutation values of the test statistic and to take into account the overall distribution of individual test statistics, the entire data set was used rather than only the genes in the gene sets tested.

Flow cytometry data processing. Gating was performed using either FlowJo version 9.7.5 or the FlowCore R package version 1.28.24. For computational modelling, values were normalized by subtracting the average values of untreated controls. For the $CD8/T_{reg}$ ratio, the percentage of $CD8^+CD44^+$ cells were divided by the percentage of $CD4^+FOXP3^+$ cells. Because these data could be skewed with varying and wide distributions, these data were log-transformed for downstream analysis.

Random forest for classification and survival analysis. Random forest for classification, regression, and survival analysis is a multivariable non-parametric ensemble partitioning tree method that can be used to model the effect of all interactions between genes on a response variable[30,31]. Each model was constructed using approximately two-thirds of randomly selected samples and cross-validated on the one-third of the samples left out of the model building process (out-of-bag samples). After many iterations, results of all models were averaged to provide unbiased estimates of predicted values, error rates, and measures of variable importance. Performance of a random forest model was measured by the misclassification error rate for classification, mean squared error for regression, and by a concordance index (one minus the error rate) for survival. For each variable, an importance score was determined, which measures the contribution of the variable to the error rate (higher scores are more predictive). When multiple response variables were modelled, as in the case of determining which treatment predicts changes in a set of immune parameters, treatment groups were converted to a design matrix and importance scores were determined for each response variable. We used the "randomForestSRCM" R package version 1.2 implementation[32] and the following parameters: 1,000 trees, node size of 1, mtry values equal to the number of variables in the model, and the Breiman–Cutler permutation method for importance score determination[9]. Gini index splitting rule was used for classification and a log-rank slitting rule was used for survival analysis. For classification, stratified sampling was used when the number of samples in each class was imbalanced. All predicted values, error rates, and importance scores were calculated using out-of-bag samples to provide unbiased estimates. To account for variance due to sample size and sampling error on the accuracy of these performance measures, bootstrapping was performed using 1,000–5,000 bootstrap iterations and the mean and standard deviation of the bootstrap distribution were determined. For presentation purposes, cut-off values for predictive variables were determined by using partial plots to estimate inflection points.

Minimal depth was used as a rigorous method to select predictive variables. Minimal depth (MD) is a dimensionless statistic that we have recently described[8] that measures the predictiveness of a variable in tree-based models. Specifically, MD measures the shortest distance from the root node of a classification/regression tree to the parent node of a maximal subtree for a variable. The maximal subtree for a variable is the largest subtree whose root node splits on the variable. Thus, smaller values for MD indicate better predictiveness. A threshold value for MD that is calculated from the tree-averaged value determines whether a variable is strongly predictive. The entire MD-based variable selection is performed using two-third of the samples (in-bag samples). An unbiased prediction error rate for a model refit with the MD-selected variables is calculated using only out-of-bag samples. Using the "randomForestSRCM" package, we applied this MD-based variable selection with the same parameters used for random forest as noted above. The tree-averaged MD threshold was used. Data were bootstrapped to provide robust estimates of MD values and prediction error rates. The frequency of bootstrap models whereby the MD values for a variable was less than the MD threshold determined how often a variable was selected as a top variable, which provides an estimate for the stability of variable selection.

TCR deep sequencing and clonotype diversity analysis. DNA from pre-treatment blood, post-treatment blood, and tumour was extracted on day 16 using the Qiagen DNA extraction protocol. Samples were sequenced by Adaptive Biotechnologies using "survey" sequencing depth for tumour and "deep" sequencing depth for blood samples. Processed data were downloaded and frequencies/counts for TCR clonotypes were examined by nucleotide sequences after non-productive reads were filtered out. The top 100 most frequent TCR clonotypes in the tumour were used to examine their frequencies in the pre- and post-treatment blood. The Shannon's diversity index[33] (DI) normalized to the number of reads (DI = $-\sum(p_i \ln p_i)/\ln n$, where n is the number of clones, p_i is the clonal frequency of the ith clone, and sigma is summed from $i = 1$ to $i = n$) was calculated for each sample. This gives a value between 0 and 1, where 0 is monoclonal and 1 is an even distribution of different clones.

Unsupervised and supervised analysis of CDR3 amino acid properties. Based on previously described methods[15,16], Atchley factors were used to reduce a linear sequence of amino acids into analysable numeric features of distinct amino acid properties. The five Atchley factors and the attributes they measure are the following. (1) PAH: accessibility, polarity, and hydrophobicity; (2) PSS: propensity for secondary structure; (3) MS: molecular size; (4) CC: codon composition, (5) EC: electrostatic charge. Each CDR3 was represented as a set of all possible contiguous amino acids of length p (p-tuple). We chose $p = 3$ based on previous published reports but examined a range of P values, which gave comparable results (see below). For each p-tuple, the Atchley factors for the amino acids were then calculated to give a vector of length $5p$, or 15 (3 amino acids × 5 Atchley factors). Thus, each CDR3 was represented by a set of these vectors. The average values for these vectors were calculated for the top B most frequent clones from the post-treatment blood. A cut-off of $B = 5$ was chosen based on examination of the frequency distribution of the TCR clonotypes and an estimate of the number of clones with extreme values compared to the rest of the distribution. These averaged values were then clustered into two groups by k-means clustering with $k = 2$. The association between cluster membership and treatment with or without radiation was calculated by Fisher's exact test. This entire process was repeated for the five clones in the pre-treatment blood, for randomly drawn clones from the post-treatment blood, for p-tuple lengths from $p = 2$ to 10, and for cut-off values from $B = 3$ to 50. In all cases, the distribution of P values was compared to the P value from the observed data.

Although averaging the Atchley factor values is a simple method to agglomerate CDR3 features for unsupervised classification, it does not provide insight into how treatment groups influence the amino acids that comprise the CDR3. To understand which sets of p-tuples were most strongly influenced by treatment groups with radiation, without radiation, and pre-treatment blood, we used previously described methods[15] to assign p-tuples into n clusters based on their Atchley factor vector. Model based clustering with cluster number determination using the "mclust" R package was applied to all p-tuples from the top five clones in all treatment groups from pre- and post-treatment blood. This gave rise to 17 clusters, or subsets, of p-tuples. The proportion of p-tuples belonging to each of these 17 subsets, denoted P_i, was then calculated for each clonotype and used as features. The subsets that were most influenced by treatment group (treatment group with radiation, without radiation, or pre-treatment) were then analysed by multivariable random forest regression using a design matrix for treatment groups as the x-variable and P_i as the response variable. The variables P_i most affected by each treatment group were selected by comparing the observed importance scores to the importance scores generated by permutation. To determine the location and frequencies of amino acids belonging to the selected p-tuple subsets across the variable length CDR3 region, the CDR3 of each clone was divided into 10 bins of equal size. Then, the proportion of p-tuples in each of these 10 bins belonging to the selected subset was calculated and compared between treatment groups.

21. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45,** 228–247 (2009).
22. Bayne, L. J. et al. Tumor-derived granulocyte-macrophage colony-stimulating factor regulates myeloid inflammation and T cell immunity in pancreatic cancer. *Cancer Cell* **21,** 822–835 (2012).
23. Lee, Y. et al. Therapeutic effects of ablative radiation on local tumor require $CD8^+$ T cells: changing strategies for cancer treatment. *Blood* **114,** 589–595 (2009).
24. Dewan, M. Z. et al. Fractionated but not single-dose radiotherapy induces an immune-mediated abscopal effect when combined with anti-CTLA-4 antibody. *Clin. Cancer Res.* **15,** 5379–5388 (2009).
25. Green, M. R. et al. Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. *Blood* **116,** 3268–3277 (2010).
26. Holets, L. M. Trophoblast CD274 (B7-H1) is differentially expressed across gestation: influence of oxygen concentration. *Biol. Reprod.* **74,** 352–358 (2006).
27. Cheadle, C., Vawter, M. P., Freed, W. J. & Becker, K. G. Analysis of microarray data using Z score transformation. *J. Mol. Diagn.* **5,** 73–81 (2003).

LETTER

Theileria parasites secrete a prolyl isomerase to maintain host leukocyte transformation

J. Marsolier[1], M. Perichon[1], J. D. DeBarry[2], B. O. Villoutreix[3], J. Chluba[4,5], T. Lopez[4,5], C. Garrido[4,5,6], X. Z. Zhou[7], K. P. Lu[7], L. Fritsch[1], S. Ait-Si-Ali[1], M. Mhadhbi[8], S. Medjkane[1]* & J. B. Weitzman[1]*

Infectious agents develop intricate mechanisms to interact with host cell pathways and hijack their genetic and epigenetic machinery to change host cell phenotypic states. Among the Apicomplexa phylum of obligate intracellular parasites, which cause veterinary and human diseases, *Theileria* is the only genus that transforms its mammalian host cells[1]. *Theileria* infection of bovine leukocytes induces proliferative and invasive phenotypes associated with activated signalling pathways, notably JNK and AP-1 (ref. 2). The transformed phenotypes are reversed by treatment with the theilericidal drug buparvaquone[3]. We used comparative genomics to identify a homologue of the peptidyl-prolyl isomerase PIN1 in *T. annulata* (TaPIN1) that is secreted into the host cell and modulates oncogenic signalling pathways. Here we show that TaPIN1 is a bona fide prolyl isomerase and that it interacts with the host ubiquitin ligase FBW7, leading to its degradation and subsequent stabilization of c-JUN, which promotes transformation. We performed *in vitro* and *in silico* analysis and *in vivo* zebrafish xenograft experiments to demonstrate that TaPIN1 is directly inhibited by the anti-parasite drug buparvaquone (and other known PIN1 inhibitors) and is mutated in a drug-resistant strain. Prolyl isomerization is thus a conserved mechanism that is important in cancer and is used by *Theileria* parasites to manipulate host oncogenic signalling.

To identify proteins secreted by *Theileria* into the host cell that could contribute to transformation[4–6], we conducted an *in silico* screen of parasite genomes; we identified 689 proteins in the *T. annulata* genome with a predicted signal peptide. Comparison with the *T. gondii* (a non-transforming apicomplexan parasite) proteome narrowed the candidate list to 33 proteins with a *Theileria*-specific signal peptide (Extended Data Fig. 1a). We focused on the *TA18945* gene encoding a homologue of the human parvulin PIN1 (hPIN1) peptidyl-prolyl isomerase (PPIase), as mammalian PIN1 regulates cell proliferation, pluripotency and survival[7,8], and contributes to tumorigenesis[9,10]. hPIN1 catalyses the *cis/trans* isomerization of peptidyl-prolyl bonds in phosphorylated Ser/Thr-Pro motifs, inducing conformational changes that affect substrate stability and activity[11,12], and there are several small-molecule inhibitors of hPIN1 (refs 13–15). The *TA18945*-encoded protein has a signal peptide and a highly conserved PPIase domain (Extended Data Fig. 1b, c), but lacks the WW domain important for substrate recognition of mammalian PIN1 (ref. 11). A gene in the *T. parva* genome, also associated with transformation, encodes a conserved TpPIN1 predicted protein, whereas the signal peptide is not conserved in the related *T. orientalis* genome, which does not transform host cells[16] (Extended Data Fig. 2a, b). We detected *Theileria pin1* transcripts in B cells infected with *T. annulata* or *T. parva*, and they decreased upon buparvaquone treatment (Fig. 1a). The levels of host bovine Bt*Pin1* transcripts were unaffected by *Theileria* infection or buparvaquone treatment (Extended Data Fig. 3). An antibody generated against a TaPIN1-specific peptide (NPVNRN

TGMAVTR) recognized parasite PIN1 protein or transfected TaPIN1 in mouse fibroblasts, but not mammalian PIN1 (Fig. 1b and Extended Data Fig. 4a–e). Confocal microscopy and immunoblot analysis located the parasite PIN1 protein to both the host cell cytoplasm and nucleus (Fig. 1b, c and Extended Data Fig. 4c, d). The host nuclear signal in the confocal images was tenfold over background in parasitized cells (205.0 ± 15.48 nuclear fluorescence intensity per pixel compared with 21.45 ± 8.50 in controls, $P < 0.0001$, $n = 31$). Thus, comparative parasite genomics identified TaPIN1, which is secreted into the host cytoplasm and nucleus.
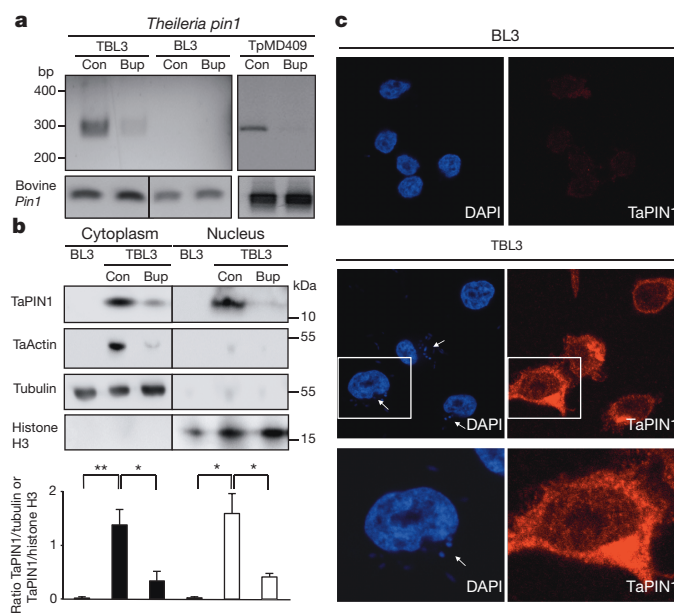
Figure 1 | *Theileria* parasites secrete a conserved PIN1 PPIase protein. **a**, Expression of *pin1* RNA in *T. annulata*-infected TBL3 cells, uninfected BL3 cells or *T. parva*-infected TpMD409 cells, treated with buparvaquone (Bup) or control (Con). Bt*Pin1* expression was used as loading control. **b**, TaPIN1 protein was detected in the host cytoplasm and nucleus, in contrast with apicomplexan actin (TaActin). Bovine histone H3 (nuclear) and tubulin (cytoplasmic) proteins were controls. Relative quantification showing TaPIN1/tubulin or TaPIN1/histone H3 ratios was calculated with Image J software (average ± standard deviation (s.d.), $n = 3$). The *P* values were corrected for multiple comparisons using the Bonferroni correction based on the total overall number of pairwise comparisons. *$P < 0.05$, **$P < 0.01$. All original western blots are shown in Extended Data Fig. 10. **c**, TaPIN1 was detected in the cytoplasm and nucleus of infected cells by confocal microscopy using an affinity-purified antibody specific for TaPIN1, counterstaining with 4′,6-diamidino-2-phenylindole (DAPI) (white arrows indicate parasites). Objective used, 60×; magnification in bottom panels, 600×. Results are representative of three independent experiments.

[1]Université Paris Diderot, Sorbonne Paris Cité, Epigenetics and Cell Fate, UMR 7216 CNRS, 75013 Paris, France. [2]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, Georgia 30602, USA. [3]Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques *in silico*, INSERM UMR-S 973, 75013 Paris, France. [4]INSERM, UMR 866, Equipe labellisée Ligue contre le Cancer and Laboratoire d'Excellence LipSTIC, 21000 Dijon, France. [5]University of Burgundy, Faculty of Medicine and Pharmacy, 21000 Dijon, France. [6]Centre anticancéreux George François Leclerc, CGFL, 21000 Dijon, France. [7]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. [8]Laboratoire de Parasitologie, Ecole Nationale de Médecine Vétérinaire, Université de la Manouba, 2020 Sidi Thabet, Tunisia.
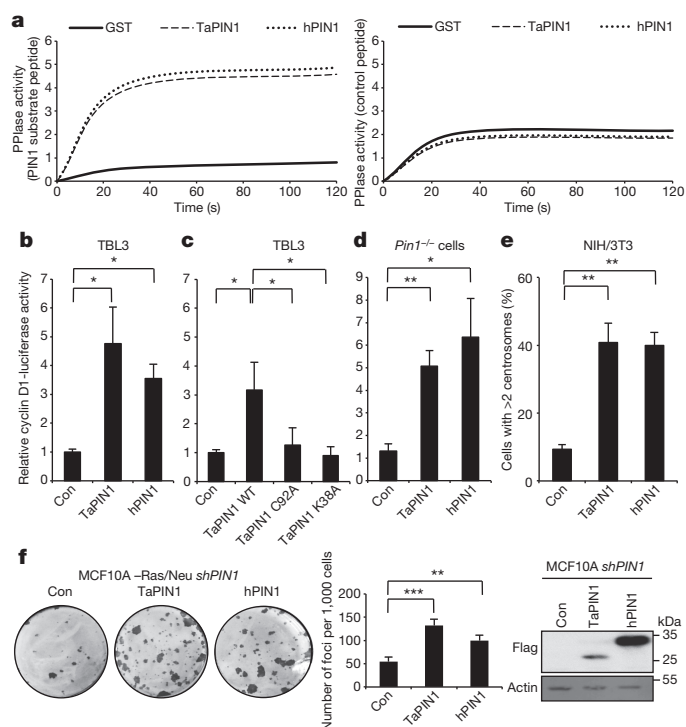*These authors contributed equally to this work.

To explore the functional PPIase activity of the secreted TaPIN1 protein, we developed a chymotrypsin-coupled *in vitro* assay and found that TaPIN1 and hPIN1 catalytic activities were comparable (Fig. 2a). TaPIN1 and hPIN1 were also equivalent in activation of the cyclin D1-luciferase reporter in bovine B cells (Fig. 2b), an established readout for PIN1 activity[9]. We mutated key C92 and K38 residues in TaPIN1 and showed loss of the PPIase activity (Fig. 2c). Furthermore, TaPIN1 rescued cyclin D1 promoter activity and cell spreading defects in $Pin1^{-/-}$ mouse fibroblasts (Fig. 2d and Extended Data Fig. 5a). Mammalian PIN1 overexpression disrupts cell cycle regulation, causing centrosome amplification and cell transformation[17]. TaPIN1 also induced centrosome duplication when overexpressed in mouse fibroblasts (Fig. 2e and Extended Data Fig. 5b). Furthermore, TaPIN1 functionally replaced mammalian PIN1 and rescued colony formation as effectively as hPIN1 in human breast cancer cells with knocked-down PIN1 (Fig. 2f). These combined results show that *Theileria* secretes a bona fide phosphorylation-dependent PPIase that could contribute to host cell transformation.
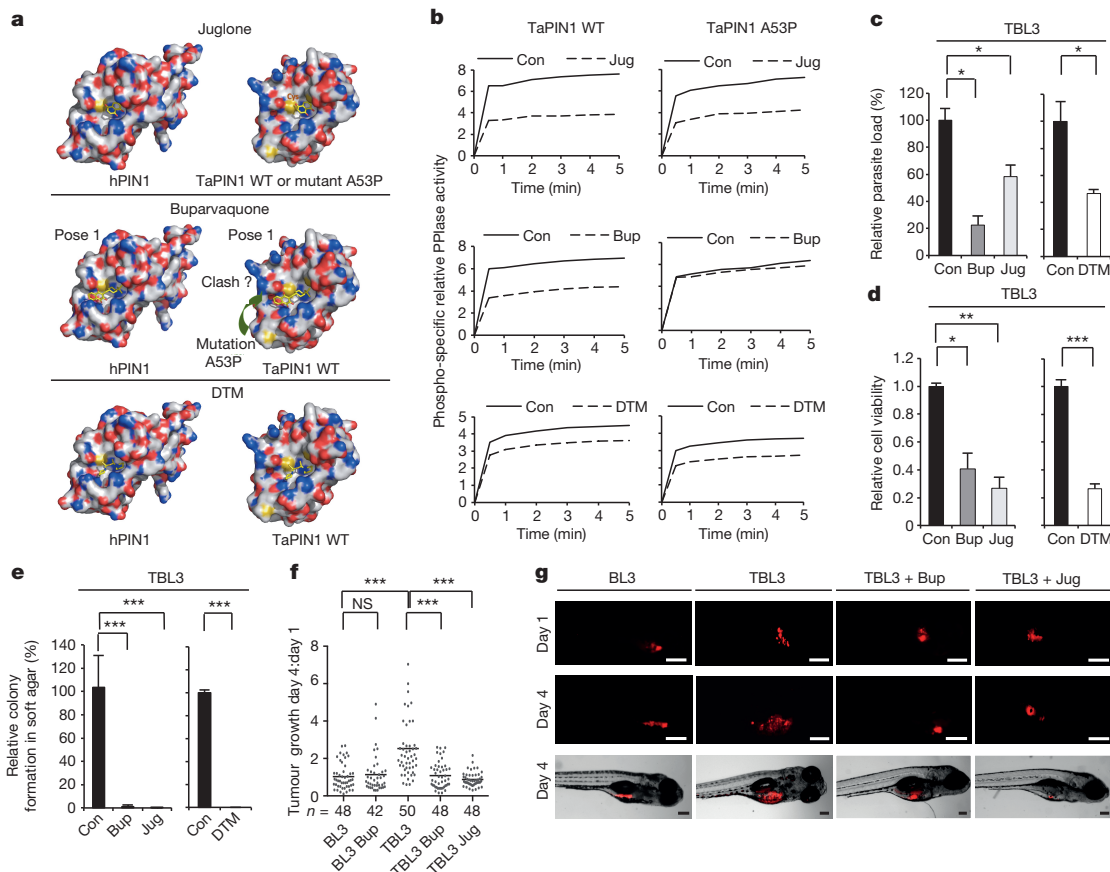
In a search for potential inhibitors, we noted that the chemical structure of buparvaquone is similar to juglone, a well characterized inhibitor

Figure 2 | TaPIN1 is a functional homologue of hPIN1 involved in transformation. a, hPIN1 and TaPIN1 catalytic PPIase activities measured by *in vitro* chymotrypsin-coupled assay using a PIN1 substrate peptide (Suc-Ala-Glu-Pro-Phe-pNA). No activity was detected for glutathione *S*-transferase (GST) alone or control substrate peptide (Suc-Ala-Ala-Pro-Phe-pNA). **b**, TaPIN1 and hPIN1 increased cyclin D1-luciferase promoter activity when transfected in TBL3 cells. **c**, C92A and K38A TaPIN1 mutants showed reduced activation of cyclin D1 promoter when transfected in TBL3 cells. WT, wild type. **d**, TaPIN1 or hPIN1 induced cyclin D1-luciferase promoter activity in $Pin1^{-/-}$ immortalized fibroblasts. **e**, TaPIN1 causes centrosome amplification. NIH/3T3 fibroblasts stably expressing TaPIN1 or hPIN1 were arrested at the G1/S transition by aphidicolin, stained with anti-γ-tubulin antibody. Three-hundred cells were scored. **f**, TaPIN1 or hPIN1 transfection increased colony foci formation in PIN1-knockdown MCF10A–Ras/Neu cells. Expression of transfected TaPIN1 and hPIN1 in MCF10A–Ras/Neu cells with short hairpin RNA (shRNA)-silenced *PIN1* detected with an anti-Flag antibody. Actin was a loading control. Con, control empty vector transfection. Data represent three independent experiments (average ± s.d.). **c**, *P* values were calculated using the Dunnett method for multiple comparisons with the TaPIN1 wild type. **b, d–f**, *P* values were corrected using Dunnett multiple comparisons with the control. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

of mammalian PIN1 (ref. 13). The TaPIN1 sequence exhibits over 47% identity with hPIN1 in the PPIase domain (Extended Data Fig. 6a). Our homology models of TaPIN1 protein, based on published hPIN1 experimental data, suggest a similar structure with a conserved catalytic pocket (Fig. 3a and Extended Data Fig. 6b). Notably, several PIN1 homologues also lack the WW domain, including *Arabidopsis thaliana* AtPIN1[18–20], MdPIN1 in *Malus domestica* and the parasite *Trypanosoma brucei* TbPIN1 homologue[20–22], and the predicted TaPIN1 model closely resembles these structures (Extended Data Fig. 6d). We investigated the hPIN1 experimental structure and the TaPIN1 predicted model with the binding pocket and hotspot detection algorithm FTMap, using the server FTFlex. Notably, we found key hotspot regions in the catalytic site area, matching the substrate-binding region of hPIN1 (Extended Data Fig. 6). Juglone and buparvaquone molecules could be docked into the active site of both TaPIN1 and hPIN1 using *in silico* approaches (Fig. 3a and Extended Data Fig. 6c). We predicted that buparvaquone might target TaPIN1 directly and that juglone (or other PIN1 inhibitors) could functionally replace buparvaquone to block parasite transformation. Both buparvaquone and juglone inhibited TaPIN1 PPIase activity *in vitro*, as did the unrelated non-quinone inhibitor dipentamethylene thiuram monosulphide (DTM)[14], albeit to a lesser degree (Fig. 3b). Buparvaquone-resistant *Theileria* strains are an emerging clinical concern for cattle in infected areas[23] and mutations in the cytochrome *b* gene were recently reported[24]. But mitochondrial and non-mitochondrial pathways might cooperate in transformation and participate in drug resistance. We sequenced the Ta*pin1* gene in genomic DNA from a drug-resistant isolate and identified a mutation (A53 > P substitution) in the catalytic loop of TaPIN1 (Extended Data Fig. 7). Structural modelling suggested that this mutation could affect the nearby catalytic region and disturb ligand binding; computational docking indicated that the small juglone molecule could react with the thiolate group of C113 or C92 in hPIN1, TaPIN1 or mutant TaPIN1(A53P). However, the A53P mutation might impede interaction with the bulky, hydrophobic moiety of the larger buparvaquone compound (molecular weight (MW) = 326 Da, compared with juglone MW = 176 Da), creating steric clashes between the inhibitor and residues in the modified structure (Fig. 3a and Extended Data Fig. 6c). Mutant TaPIN1(A53P) was catalytically active on the PIN1 substrate and was inhibited by juglone and DTM, but not by buparvaquone (Fig. 3b). The PIN1 inhibitors (buparvaquone, juglone and DTM) all reduced parasite load and viability of host cells infected with *T. annulata* or *T. parva* (Fig. 3c, d and Extended Data Fig. 8a) and blocked colony growth of parasitized cells in soft-agar assays *in vitro* (Fig. 3e and Extended Data Fig. 8b). In contrast, knocking down the endogenous bovine Bt*Pin1* did not affect colony formation (Extended Data Fig. 8c). Transfection with mutant TaPIN1(A53P) rendered TBL3 cells resistant to buparvaquone, but not juglone, treatment (Extended Data Fig. 8d). Similarly, juglone inhibited both wild-type and mutant TaPIN1 activity in the cyclin D1-luciferase assay, but only the mutant was resistant to buparvaquone (Extended Data Fig. 8e). Fish xenograft models are effective for monitoring *in vivo* tumour formation and for drug testing[25], and are emerging as important experimental models to study cancer[26]. We used a zebrafish xenograft experimental system to test drug effects on tumour growth *in vivo* and observed a twofold increase in tumour growth of infected cells that was efficiently inhibited by the anti-PIN1 drugs (Fig. 3f, g). Thus, our modelling predictions, biochemical analysis *in vitro*, transformation assays and tumour growth *in vivo* all support the targeting of TaPIN1 by buparvaquone and the role of TaPIN1 in *Theileria*-induced cell transformation.

To investigate how TaPIN1 affects host signalling pathways, we studied relevant substrates targeted by TaPIN1. hPIN1 targets many proteins, including the ubiquitin ligase FBW7 (ref. 27), which exerts an anti-tumour function by degrading oncoproteins required for cellular proliferation, such as c-JUN[28,29]. Since c-JUN is induced (and critical) during *Theileria*-induced transformation[2], we examined whether TaPIN1 targets the conserved bovine FBW7 protein. We found that TaPIN1 interacts

Figure 3 | Inhibition of TaPIN1 activity blocks transformation *in vitro* and *in vivo*. a, Homology prediction models for TaPIN1 and TaPIN1(A53P) mutant based on similarity with hPIN1. The TaPIN1 A53P mutation induces a conformational change near the catalytic loop (green arrow). Computational analysis predicted docking of juglone, buparvaquone or DTM molecules in the PIN1 active sites (see Extended Data Fig. 6c for alternative). Buparvaquone/ juglone: red, polar and negatively charged residues; blue, polar and positively charged residues; yellow, S atoms; white, remaining residues. DTM: blue, N atoms; amber, S atoms. Colours in small molecule: O atoms in red; other atoms in yellow. WT, wild type. **b**, TaPIN1 and TaPIN1(A53P) catalytic PPIase activity, measured with an *in vitro* chymotrypsin-coupled assay, upon treatment with buparvaquone (Bup), juglone (Jug) or DTM. Con, control solutions. **c**, Drug treatment (72 h) eliminated *Theileria* parasites in infected cells (parasite nuclei were counted after DAPI staining). **d**, PIN1 inhibitors decreased the viability of infected TBL3 cells (XTT assay 72 h). **e**, Drug treatment decreased colony formation of parasitized cells in soft agar (72 h treatment with buparvaquone, juglone or DTM: macroscopic colonies per plate after 10 days). **f**, TaPIN1 inhibition reduced xenograft tumour growth in zebrafish embryos (with a Zeiss AxioZoom V16 Macroscope at day 1 and day 4 post-injection). The median tumour development day 4:day 1 ratio is shown. *n*, number of embryos. **g**, Representative images from individual zebrafish embryos photographed with a Zeiss AxioZoom V16 Macroscope. Scale bars, 200 μm. Con, vector solutions alone. All data represent three independent experiments (average ± s.d., *n* = 3). **c–e**, *P* values were corrected using the Dunnett test multiple comparisons with the control. **f**, An unpaired Mann–Whitney test was performed for the zebrafish experiments to analyse the significant differences between the control and treatment groups. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. NS, not significant.

with host FBW7 in *Theileria*-infected cells or murine fibroblasts (Extended Data Fig. 9a, b). Conversely, FBW7α in particular, but not other isoforms, co-immunoprecipitated TaPIN1 from parasitized cells (Fig. 4a). FBW7α protein levels were reduced in parasitized cells compared with uninfected cells and correlated with elevated c-JUN levels (Fig. 4b). Pharmacological TaPIN1 inhibition restored FBW7 protein expression and reduced c-JUN levels, without affecting messenger RNA expression (Fig. 4b and Extended Data Fig. 9c, d). But knocking down BtPIN1 did not affect FBW7 or c-JUN protein levels (Extended Data Fig. 9e). Short interfering RNA (siRNA) knockdown of FBW7 caused accumulation of c-JUN protein, whereas exogenous FBW7α transfection decreased c-JUN protein levels (Fig. 4c). Furthermore, knocking down FBW7 increased AP-1 activity, as measured by a luciferase reporter assay, and rescued the inhibition of AP-1 by buparvaquone or juglone (Fig. 4d, e). TaPIN1 inhibition caused increased c-JUN ubiquitination in TBL3 cells and decreased FBW7 (auto-)ubiquitination (Extended Data Fig. 9f). c-JUN ubiquitination was FBW7-dependent, as this effect was abolished by siRNA targeting bovine *Fbw7* (Fig. 4f). Half-life analysis using cyclo-heximide showed that *siFbw7* increased c-JUN stability and rescued c-JUN levels after TaPIN1 inhibition (Fig. 4g). In addition, the TaPIN1

(A53P) mutant rescued the effect of buparvaquone, but not juglone, on c-JUN ubiquitination and transcriptional activity (reflected by the expression of bovine *Mmp-9*, an AP-1 target gene) (Extended Data Figs 8f and 9g). As mammalian FBW7 targets many protein substrates, we examined the effects of buparvaquone treatment or FBW7α trans-fection, but noted no changes in levels of endogenous c-MYC or acti-vated NOTCH1 proteins, while there was a modest effect on the KLF5 transcription factor (Extended Data Fig. 9h, i). These combined results suggest that c-JUN is the major target of TaPIN1–FBW7α in parasit-ized cells. Finally, FBW7α overexpression or c-JUN knockdown both caused a significant reduction of colony growth in soft-agar prolifera-tion assays of parasitized TBL3 cells (Fig. 4h).

c-JUN is critical for *Theileria* transformation and host cell prolifera-tion[2], but it was previously unknown how the parasite initiated this effect. In TBL3 cells, c-JUN seems to be activated by reduced FBW7 de-gradation rather than phosphorylation by JNK signalling[2]. Subsequent activation of a feedback loop, involving c-JUN control of the microRNA miR-155 oncomiR, could create a epigenetic switch to maintain trans-formation and proliferation[30]. We studied B lymphocytes infected nat-urally with *T. parva* (TpMD409) or artificially with *T. annulata* (TBL3)
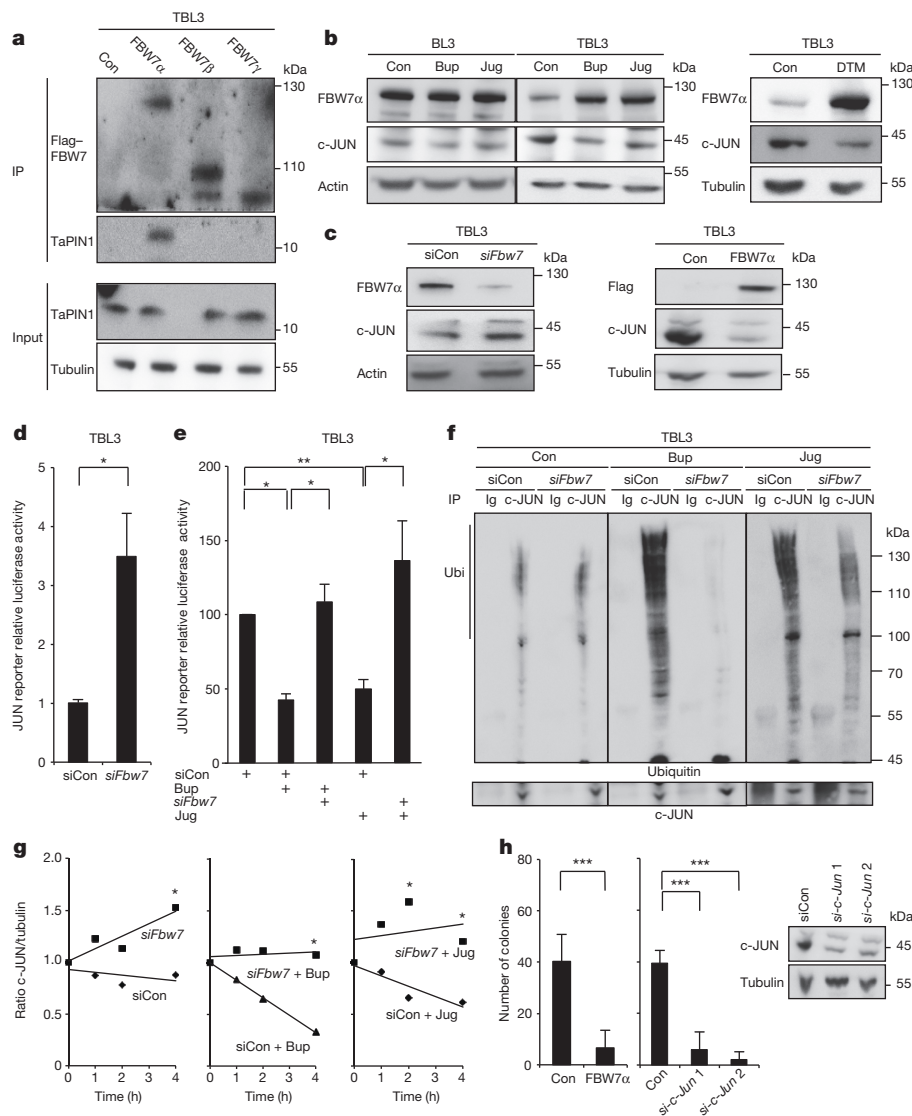
Figure 4 | TaPIN1 activates the oncogenic c-JUN pathway via FBW7 ubiquitination.
a, Endogenous TaPIN1 interacts with FBW7α isoform. Protein extracts from TBL3 cells expressing Flag–hFBW7 isoforms or Flag–control (Con) were immunoprecipitated (IP) and immunoblotted with TaPIN1 or Flag antibodies. **b**, Inhibition of TaPIN1 by buparvaquone (Bup), juglone (Jug) or DTM increased FBW7α protein levels and decreased c-JUN expression in TBL3 cells. Actin/tubulin were loading controls. **c**, Inhibition of *Fbw7* increases c-JUN protein levels in TBL3 cells, whereas ectopic FBW7α expression reduced c-JUN protein levels. Bovine actin/tubulin were loading controls. Con, empty vector. **d**, Inhibition of c-JUN reporter activity (*BIC* promoter-luciferase, an AP-1 target gene) in TBL3 cells transfected with siRNA against *Fbw7* or siControl (siCon). **e**, c-JUN reporter activity (*BIC* promoter-luciferase) in TBL3 cells treated with buparvaquone or juglone was rescued by siRNA against *Fbw7* but not siControl. **f**, *siFbw7* inhibition rescued drug effects on c-JUN ubiquitination (Ubi) in parasitized TBL3 cells incubated with MG132, followed by immunoprecipitation of endogenous c-JUN and immunoblotting with c-JUN and ubiquitin antibodies. Ig, non-specific control immunoglobulin. **g**, *siFbw7* depletion increased the half-life of endogenous c-JUN protein. TBL3 cells treated with buparvaquone or juglone were incubated with cycloheximide, followed by immunoblotting with c-JUN or tubulin antibodies. Relative c-JUN protein levels at time 0 h were set at 1. **h**, Colony formation of TBL3 cells was markedly reduced by ectopic expression of FBW7α isoform or si-c-Jun. Efficiencies of two independent *c-Jun*-targeting siRNAs are shown. Bovine tubulin loading control. All data are representative of three independent experiments (average ± s.d., n = 3). **e**, P values were calculated with the Bonferroni method based on the number of pairwise comparisons. **h**, The statistics were calculated using the Dunnett procedure. *P < 0.05, **P < 0.01, ***P < 0.001.

and are currently investigating *Theileria*-transformed T lymphocytes and macrophages. Our discovery of mutation in the Ta*pin1* gene suggests that new anti-PIN1 compounds might be effective clinical reagents to treat drug-resistant theileriosis. Finally, the evolution of a PIN1 homologue with an acquired signal peptide only in *T. annulata* and *T. parva* genomes provides a fascinating insight into how apicomplexan species have hijacked oncogenic pathways to maintain host cell transformation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 September 2013; accepted 5 November 2014.
Published online 26 January 2015.

1. Dobbelaere, D. & Heussler, V. Transformation of leukocytes by *Theileria parva* and *T. annulata*. *Annu. Rev. Microbiol.* **53**, 1–42 (1999).
2. Chaussepied, M. et al. Upregulation of Jun and Fos family members and permanent JNK activity lead to constitutive AP-1 activation in *Theileria*-transformed leukocytes. *Mol. Biochem. Parasitol.* **94**, 215–226 (1998).
3. McHardy, N., Wekesa, L. S., Hudson, A. T. & Randall, A. W. Antitheilerial activity of BW720C (buparvaquone): a comparison with parvaquone. *Res. Vet. Sci.* **39**, 29–33 (1985).
4. Shiels, B. R. et al. A *Theileria annulata* DNA binding protein localized to the host cell nucleus alters the phenotype of a bovine macrophage cell line. *Eukaryot. Cell* **3**, 495–505 (2004).
5. Pain, A. et al. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva. Science* **309**, 131–133 (2005).
6. Witschi, M. et al. Proteomic analysis of the *Theileria annulata* schizont. *Int. J. Parasitol.* **43**, 173–180 (2013).
7. Lu, K. P., Hanes, S. D. & Hunter, T. A human peptidyl-prolyl isomerase essential for regulation of mitosis. *Nature* **380**, 544–547 (1996).
8. Winkler, K. E., Swenson, K. I., Kornbluth, S. & Means, A. R. Requirement of the prolyl isomerase Pin1 for the replication checkpoint. *Science* **287**, 1644–1647 (2000).
9. Wulf, G. M. et al. Pin1 is overexpressed in breast cancer and cooperates with Ras signaling in increasing the transcriptional activity of c-Jun towards cyclin D1. *EMBO J.* **20**, 3459–3472 (2001).
10. Ryo, A. et al. PIN1 is an E2F target gene essential for *Neu/Ras*-induced transformation of mammary epithelial cells. *Mol. Cell. Biol.* **22**, 5281–5295 (2002).
11. Yaffe, M. B. et al. Sequence-specific and phosphorylation-dependent proline isomerization: a potential mitotic regulatory mechanism. *Science* **278**, 1957–1960 (1997).
12. Lu, K. P., Finn, G., Lee, T. H. & Nicholson, L. K. Prolyl *cis-trans* isomerization as a molecular timer. *Nature Chem. Biol.* **3**, 619–629 (2007).
13. Hennig, L. et al. Selective Inactivation of parvulin-like peptidyl-prolyl *cis/trans* isomerases by juglone. *Biochemistry* **37**, 5953–5960 (1998).
14. Tatara, Y., Lin, Y.-C., Bamba, Y., Mori, T. & Uchida, T. Dipentamethylene thiuram monosulfide is a novel inhibitor of Pin1. *Biochem. Biophys. Res. Commun.* **384**, 394–398 (2009).
15. Moore, J. D. & Potter, A. Pin1 inhibitors: pitfalls, progress and cellular pharmacology. *Bioorg. Med. Chem. Lett.* **23**, 4283–4291 (2013).
16. Hayashida, K. et al. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of *Theileria*-induced leukocyte transformation. *MBio* **3**, e00204–12 (2012).
17. Chen, C.-H. et al. SENP1 desumoylates and regulates Pin1 protein activity and cellular function. *Cancer Res.* **73**, 3951–3962 (2013).
18. Landrieu, I. et al. The *Arabidopsis thaliana PIN1At* gene encodes a single-domain phosphorylation-dependent peptidyl prolyl *cis/trans* isomerase. *J. Biol. Chem.* **275**, 10577–10581 (2000).

19. Landrieu, I., Wieruszeski, J.-M., Wintjens, R., Inzé, D. & Lippens, G. Solution structure of the single-domain prolyl *cis/trans* isomerase PIN1At from *Arabidopsis thaliana*. *J. Mol. Biol.* **320,** 321–332 (2002).

20. Yao, J.-L., Kops, O., Lu, P.-J. & Lu, K. P. Functional conservation of phosphorylation-specific prolyl isomerases in plants. *J. Biol. Chem.* **276,** 13517–13523 (2001).

21. Goh, J. Y. *et al.* Functional characterization of two novel parvulins in *Trypanosoma brucei*. *FEBS Lett.* **584,** 2901–2908 (2010).

22. Sun, L. *et al.* Solution structural analysis of the single-domain parvulin TbPin1. *PLoS ONE* **7,** e43017 (2012).

23. Mhadhbi, M. *et al. In vivo* evidence for the resistance of *Theileria annulata* to buparvaquone. *Vet. Parasitol.* **169,** 241–247 (2010).

24. Sharifiyazdi, H., Namazi, F., Oryan, A., Shahriari, R. & Razavi, M. Point mutations in the *Theileria annulata* cytochrome *b* gene is associated with buparvaquone treatment failure. *Vet. Parasitol.* **187,** 431–435 (2012).

25. Konantz, M. *et al.* Zebrafish xenografts as a tool for *in vivo* studies on human cancer. *Ann. NY Acad. Sci.* **1266,** 124–137 (2012).

26. White, R., Rose, K. & Zon, L. Zebrafish cancer: the state of the art and the path forward. *Nature Rev. Cancer* **13,** 624–636 (2013).

27. Min, S.-H. *et al.* Negative regulation of the stability and tumor suppressor function of Fbw7 by the Pin1 prolyl isomerase. *Mol. Cell* **46,** 771–783 (2012).

28. Nateri, A. S., Riera-Sans, L., Costa, C. D. & Behrens, A. The ubiquitin ligase SCFFbw7 antagonizes apoptotic JNK signaling. *Science* **303,** 1374–1378 (2004).

29. Wei, W., Jin, J., Schlisio, S., Harper, J. W. & Kaelin, W. G. The v-Jun point mutation allows c-Jun to escape GSK3-dependent recognition and destruction by the Fbw7 ubiquitin ligase. *Cancer Cell* **8,** 25–33 (2005).

30. Marsolier, J. *et al.* OncomiR addiction is generated by a miR-155 feedback loop in *Theileria*-transformed leukocytes. *PLoS Pathog.* **9,** e1003222 (2013).

Author Contributions J.M. and M.P. performed the experiments; J.D.D. performed the comparative genomics bioinformatics screen; B.O.V. performed the structural three-dimensional modelling; J.C., T.L. and C.G. designed and executed the zebrafish experiments; L.F., S.A.-S.-A., X.Z.Z. and K.P.L. provided critical reagents and advice; M.M. provided genomic DNA from buparvaquone-resistant parasites; J.M., S.M. and J.B.W. conceived the study and designed experiments, analysed the data and wrote the paper. All authors read the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.B.W. (jonathan.weitzman@univ-paris-diderot.fr).

METHODS

Cell lines and culture conditions. All infected bovine cell lines used in this study were previously described: TBL3 cells were derived from *in vitro* infection of the spontaneous bovine B lymphosarcoma cell line, BL3, with Hissar stock of *T. annulata*. The TpMD409 lymphocyte cell line is infected with *T. parva*. The culture conditions of these cell lines were described previously[31]. All parasite-infected cell lines were provided by the Langsley laboratory. Cells were cultured in RPMI 1640 (Gibco-BRL), supplemented with 10% heat-inactivated fetal calf serum, 4 mM L-glutamine, 25 mM HEPES, 10 μM β-mercaptoethanol and 100 μg ml^{-1} penicillin/streptomycin in a humidified 5% CO_2 atmosphere at 37 °C. NIH/3T3 and murine immortalized fibroblast cells were provided by C. Francastel and G. Del Sal, respectively. MCF10A–Ras/Neu *shPIN1* cells were previously described[17]. Murine and human cell lines were cultured in DMEM (Gibco-BRL), supplemented with 10% heat-inactivated fetal calf serum, 4 mM L-glutamine and 100 μg ml^{-1} penicillin/streptomycin in a humidified 5% CO_2 atmosphere at 37 °C. Cell numbers, as judged by Trypan Blue exclusion test, were determined by counting cells using a Countess automated cell counter (Invitrogen). All cell lines were mycoplasma negative. The anti-parasite drug buparvaquone[3] (BW720c) was used at 200 ng ml^{-1} for 72 h (Chemos GmbH, ref: 88426-33-9). BW720c has no effect on growth of uninfected cells (Hudson, 1985). Cells were treated with juglone at 5 μM resuspended in ethanol (Sigma, ref: H47003), DTM at 1 μM resuspended in dimethylsulphoxide (DMSO) (DTM was provided by T. Uchida).

Plasmids. Plasmids p3×Flag-myc-CMV-24: FBW7α, β or γ were provided by B. E. Clurman. Human gene h*PIN1* and parasite genes Ta*PIN1* wild type (*TA18945*) or TaCyclophilin (*TA19600*) were cloned between restriction sites XhoI and NotI in pREV-HA-Flag-RIL2 using oligonucleotides: h*PIN1* forward, CCGCTCGAG GCGGACGAGGAGAAGCTG, reverse, AAGGAAAAAAGCGGCCGCTCACT CAGTGCGGAGGATGA; Ta*pin1* wild type, forward, CCGCTCGAGGCCCACT TGCTACTAAAG, reverse, ATAAGAATGCGGCCGCTTATGCGATTCTATA TATAAGATG; and TaCyclophilin, forward, CCGCTCGAGTTCTACAATCAA CCCAAGCAT, reverse, AAGGAAAAAAGCGGCCGCTCACAATAATTCTCC ACAGTCC. Point mutations TaPIN1 K38A, A53P and C92A were created from pRev-HA-Flag-TaPin1 WT-RIL2 using a set of primers following a three-step PCR protocol: Ta*pin1* K38A forward, GCCCACTTGCTACTAGCGCACACTGGATC TAGG, reverse, CCTAGATCCAGTGTGCGTAGTAGCAAGTGGGC; Ta*pin1* A53P, forward, GGAATACTGGAATGCCAGTAACAAGAAC, reverse, GTTC TTGTTACTGGCATTCCAGTATTCC; and Ta*pin1* C92A, forward, GCAACTG CCAAATCTGAGGCTTCAAGCGCAAGAAAAGG, reverse, CCTTTTCTTGC GCTTGAAGCCTCAGATTTGGCAGTTGC.

siRNA. BL3 and TBL3 cells were transfected using Neon Transfection kit (Invitrogen). Cells were double transfected with 400 nM of the indicated siRNA: *siFbw7*, CATCATTAGTGGATCCACGG; *siPin1*, GCCATTTGAAGACGCCTCC; *si-c-Jun* 1, CCACGGCCAAUAUGCUCAGG; *si-c-Jun* 2, AUGACUGCAAAGAUGGAAA.

Parasite genomic DNA extraction and sequencing. Buparvaquone-resistant infected cells were cultured in RPMI 1640 and parasite DNA was extracted using the kit Promega (Wizard Genomic DNA Purification Kit, ref: A1125) following the manufacturer's instructions. To sequence the Ta*pin1* gene, we first performed a PCR with specific oligonucleotides (forward, GTCTGTCAAATAGGTAGAAA TC, reverse, GAGAGGAAGTTGAATCAAACAT) using High Fidelity Platinium Taq Polymerase (Invitrogen, ref: 11304) and sequencing was performed using the same oligonucleotides.

RNA extraction and RT–qPCR. Total cellular RNAs were extracted using a NucleoSpin RNA Kit (Macherey Nagel, ref: 740955) and parasite RNAs were extracted using the classical Trizol Protocol. cDNA synthesis was performed with the Reverse Transcriptase Superscript III (Invitrogen, Ref: 18080051). Quantitative PCR amplification was performed using the Sybr Green reagent (Applied Biosystems, ref: 4309155). *JUN*, forward, ACGTTTTGAGGCGAGACTGT, reverse, TCTGTTTC CCTCTCGCAACT; *FBW7*, forward, AGCTGGAGTGGACCAGAGAAATTG, reverse, GAATGAGAGCACGTAAAGTGC, *PIN1*, forward, GGCCGGGTGTA CTACTTCAA, reverse, TTGGTTCGGGTGATCTTCTC; *MMP9*, forward, CCC ATTAGCACGCACGACAT, reverse, TCACGTAGCCCACATAGTCCA; *H2A*, forward, GTCGTGGCAAGCAAGGAG, reverse GATCCGGCCGTTAGGTAC TC; and β-actin, forward, GGCATCCTGACCCTCAAGTA, reverse, CACACG GAGCTCGTTGTAGA. The detection of a single product was verified by dissociation curve analysis. Relative quantities of mRNA were analysed using the ΔC_t method. The β-actin and *H2A* qPCR were used for normalization.

Nucleus/cytoplasmic protein extraction. Cells were lysed in the following buffer: 5 mM Tris HCl pH 7.5, 40 mM KCl, 2 mM $MgCl_2$, 0.5 mM EDTA, 0.05 mM spermidin and spermin, 0.1% NP-40, H_2O. Lysates were incubated for 5 min on ice and centrifuged for 10 min at 5,000 r.p.m. The supernatant constitutes the cytoplasmic fraction and the pellet constitutes the nuclear fraction.

Anti-TaPIN1 antibody purification. Anti-PIN1 rabbit polyclonal antibody raised against TaPIN1 peptide 'NPVNRNTGMAVTR' was prepared and purified by ProteoGenix SAS (Schiltigheim). Antiserum was obtained by immunizing rabbits with keyhole limpet haemocyanin (KLH)-conjugated peptide. The resulting IgG fraction was purified from antiserum by affinity chromatography against the TaPIN1 peptide.

Immunoblot analysis and immunostaining. Total proteins were extracted with Laemmli lysis buffer, sonicated: 30 s ON/30 s OFF for 5 min, resolved on 10.5% acrylamide/bis-acrylamide SDS–PAGE gels and transferred to nitrocellulose membranes (Thermo Fisher Scientific) in transfer buffer. Protein transfer was assessed by Ponceau-red staining. Membranes were blocked in Tris-buffered saline pH 7.4 containing 0.1% Tween-20 and 5% milk for 1 h at room temperature. Incubations with primary antibodies were carried out at 4 °C overnight using antibody dilutions as per the manufacturer recommendations in Tris-buffered saline pH 7.4, 0.05% Tween-20 and 5% milk. After 1 h incubation with an anti-rabbit or anti-mouse peroxidase-conjugated antibody (Jackson ImmunoResearch, ref: 111-035-003 or 115-035-003) at room temperature, proteins were detected by chemiluminescence (Thermo Fisher Scientific) following the manufacturer's instructions. We used these antibodies: rabbit anti-TaPIN1 (homemade antibody, Proteogenix, see earlier), rabbit anti-PIN1 (Cell Signalling, ref: 3722), rabbit anti-TaActin (provided by J. Baum), rabbit anti-c-JUN (Santa Cruz, ref: sc1694), mouse anti-α-tubulin (Sigma, ref: T9026), mouse anti-ubiquitin (P4D1) (Santa Cruz, ref: sc-8017), mouse anti-c-Myc (Santa Cruz, ref: sc-40), mouse anti-GST (Pierce Biotechnology, ref: MA4-004), rabbit anti-KLF5 (Abcam, ref: ab24331), rabbit anti-activated NOTCH1 (Abcam, ref: ab8925), mouse anti-HA (Roche, ref: 11583816001), rabbit anti-FBW7 (Bethyl Laboratories, ref: A301-721A), rabbit anti-histone H3 (Abcam, ref: ab1791), mouse anti-actin (Sigma, ref: A1978) and monoclonal anti-Flag M2-Peroxidase (Sigma, ref: A8592).

Parasite quantification. After indicated treatments, parasite-infected cells were plated on slides using CytoSpin centrifugation at 2,000 r.p.m. for 10 min. Cells were fixed in PBS 3.7% formaldehyde for 15 min at room temperature. Slides were mounted and coverslipped with ProLong Gold Antifade Reagent with DAPI (Invitrogen, ref: P-36931). Images of immunofluoresence staining were photographed with a fluorescent microscope (Leica Inverted 6000) and the number of parasites per cells was counted. Staining was repeated for three independent biological replicates.

PPIase assay. The PPIase activity of GST constructs: GST–control, GST–hPIN1, GST–TaPIN1 wild type and GST–TaPIN1(A53P) were determined using the protease-free PPIase activity assay[11]. The sample buffer was 35 mM HEPES (pH 7.5). We prepared stock solutions of the substrates (3 mg ml^{-1}), Suc-Ala-Glu-Pro-Phe-pNA (PIN1 substrate peptide), or a control peptide Suc-Ala-Ala-Pro-Phe-pNA (Bachem, ref: L-1635 or L-1400) in 0.47 M LiCl/TFE (anhydrous). Stock solution of chymotrypsin (100 mg ml^{-1}; Sigma, ref: C4129) was prepared in 35 mM HEPES (pH 7.8). We measured the PPIase activity with the substrate (0.03 mg ml^{-1}, 50 μM) in the presence of chymotrypsin (0.2 mg ml^{-1}) and GST–PPIases (25 nM) (pre-incubated or not with buparvaquone, juglone or DTM for 4 h at 4 °C) at 390 and 510 nm using a Flexstation III spectrophotometer.

Implantation of cells in zebrafish embryos. The zebrafish experiments described in the present study were conducted at the University of Burgundy according to French and European Union guidelines for the handling of laboratory animals. The animal procedures carried out in this study were reviewed and approved by the local Ethics Committee "Comité d'éthique de l'expérimentation animale Grand Campus Dijon" (C2EA Grand Campus Dijon number 105). Adult wild-type (WIK, ZIRC, Oregon) zebrafish and embryos were raised, staged and maintained according to standard procedures at 28 °C under a 14h: 10h light/dark cycle. Dechorionized 2 days post-fertilization zebrafish embryos were anaesthetized with 0.003% tricain (Sigma) and positioned on a 10 cm Petri dish before implantation. TBL3 or BL3 cells were treated for 24 h with DTM, buparvaquone or juglone, rinsed with PBS, labelled with the fluorescent cell tracker CM-DiI (Invitrogen) according to the manufacturer's instructions and resuspended in PBS. The cell suspensions were loaded into borosilicate glass capillary needles and the injections were performed using a microinjector (Femtojet, Eppendorf). Twenty to one-hundred cells, manually counted in injection droplets, were injected in the yolk within 3–4 h after labelling. Around 30–100 embryos were implanted per cell line. After implantation, zebrafish embryos (including non-implanted controls) were maintained at 34 °C in egg water containing 0.003% phenylthiourea (PTU). For individual tumour development analysis, each xenografted embryo was grown in a separate well in 12-well plates. Tumour growth was monitored at day 1 and and day 4 after injection by imaging the zebrafish embryos with a Zeiss AxioZoom V16 Macroscope. Images were acquired using ×2.3 objective and analysed with Zen software. For the estimation of tumour foci size, red fluorescent area was measured with Zen software and data were transferred to Excel for further calculations. No method of randomization was used to determine how animals were allocated to experimental groups.

Immunofluorescence. BL3 and TBL3 cells were plated on fibronectin-coated slides (Sigma; ref: F1141). NIH/3T3 cells and mouse immortalized fibroblast cells transfected by indicated constructs were plated on slides. All cells were then fixed in PBS

3.7% formaldehyde for 15 min at room temperature. Slides with bovine cells or NIH/3T3 cells were rinsed in PBS and permeabilized with PBS 0.2% Triton X-100 for 5 min and then blocked for 30 min with PBS 1% SVF and 1% BSA to prevent non-specific staining. These slides were incubated with rabbit anti-TaPIN1 (1/250) and/or mouse anti-HA (1/1,000; Roche, ref: 11583816001) in PBS 1% SVF and 1% BSA at room temperature for 40 min. After washing in PBS 0.2% Tween, the slides were incubated with Texas Red dye-conjugated AffinyPure donkey anti-rabbit IgG and/or Cy2 AffinyPure donkey anti-mouse IgG (1/5,000; Jackson Immunology, ref: 711-075-152 or 715-225-150) for 30 min. Slides with murine immortalized fibroblast cells were incubated for 15 min with Phalloidin-TRITC (Life Technologies, ref: R415). All slides were subsequently washed in PBS 0.2% Tween, mounted on slides and covered with ProLong Gold Antifade Reagent with DAPI (Invitrogen, ref: P-36931). Images of immunofluorescence staining (mouse immortalized fibroblast cells) were photographed with a fluorescent microscope (Leica Inverted 6000). Staining was repeated for three independent biological replicates.

Confocal microscopy analysis. Acquisitions were made on a ZEISS LSM710 laser scanning confocal. Texas Red was acquired using a 561 nm DPSS laser diode, emission captured between 587 and 690 nm. DAPI was acquired using a 405 nm laser diode, emission captured between 410 and 506 nm. Images were taken with a ×60/NA 1.4 objective, with a 2.5 zoom factor so that image pixel size was about 100 nm. Optical sections were acquired every 320 nm. Image analysis and nucleus fluorescence intensity per pixel quantifications were performed using the software Imaris 6.7.5 (Bitplane). Quantifications were done on the whole nucleus of $n = 31$ bovine infected or non-infected cells after three-dimensional construction and normalized to the background signal obtained after staining with the anti-rabbit secondary antibody alone.

Analysis of centrosome duplication during S phase. Centrosome duplication assays in NIH/3T3 cells were conducted as described previously[17]. Cells were arrested in G1/S phase by adding Aphidicolin (Sigma, ref: A0781) at a final concentration of $10\,\mu g\,ml^{-1}$ for 24 h. Cells were fixed with cold methanol for 10 min at $-20\,^{\circ}C$, then stained for centrosome with anti-γ-tubulin antibody (Sigma, ref: Clone GTU-88, T5326), and analysed by fluorescent microscopy, as described earlier.

Luciferase assay. Non-treated or treated bovine cells were transfected with the cyclin D1 or BIC luciferase reporters, using electroporation (Neon kit; Invitrogen, ref: MPK1096). Mouse cells were transfected using Lipofectamine 2000 (Invitrogen, ref: 11668019). Transfection efficiencies were normalized to Renilla activity by co-transfection of a pRL-TK Renilla reporter plasmid (Promega, ref: E6241). Luciferase assays were performed 36 h post-transfection using the Dual-Luciferase Reporter Assay System (Promega, ref: E1980) in a microplate luminometer. Relative luminescence was represented as the ratio firefly/Renilla luminescence, compared with the corresponding empty vector control.

Colony forming assay. MCF10A cells were transfected by the indicated plasmids using Fugene HD transfection system (Promega, ref: E2311) following the manufacturer's instructions. After 36 h, 1,000 cells were plated in 6-well plates. Cultures were incubated in humidified 37 °C incubators with an atmosphere of 5% CO_2 in air, and control plates were monitored for growth using a microscope. At the time of maximum foci formation (8–10 days in culture), final foci numbers were counted manually after fixation a staining with 0.5% Crystal Violet (Sigma, ref: C3886).

Soft-agar colony forming assay. A two-layer soft-agar culture system was used. A total of 20,000 bovine cells (treated with buparvaquone or juglone) or 40,000 bovine cells (treated with DTM or transfected with indicated plasmids/siRNA) were plated in a volume of 1.5 ml (0.7% SeaKem ME Agarose; Lonza, ref: 50011) plus 2× DMEM 20% fetal calf Serum over a 1.5-ml base layer (1% SeaKem ME Agarose plus 2× DMEM 20% fetal calf Serum) in 6-well plates. Cultures were incubated in humidified 37 °C incubators with an atmosphere of 5% CO_2 in air, and control plates were monitored for growth using a microscope. At the time of maximum colony formation (10–15 days in culture), final colony numbers were counted manually after fixation a staining with 0.005% Crystal Violet (Sigma, ref: C3886).

GST pull-down. hPIN1, TaPIN1 wild type and TaPIN1(A53P) were cloned between restriction sites BamHI and EcoRI in pGEX-2T plasmid, which was provided by G. Del Sal. TaPIN1 wild type or A53P: forward, CGCGGATCCGCCCACTTGCTACTAAAG, reverse, CCGGAATTCTTATGCGATTCTATATATAAGATG. Plasmid constructs were expressed in E. coli strain BL21 and purified using glutathione-sepharose beads. Concentration of purified protein was estimated by Coomassie staining. Beads coated with 1 μg of GST fusion proteins were incubated with 250 μl of cell lysate (see later) in 50 mM Tris pH 7.6, 150 mM NaCl, 0.1% Triton, for 2 h at 4 °C. Beads were washed five times with 50 mM Tris pH 7.6, 300 mM NaCl, 0.5% Triton. Proteins were revealed by western blot analysis using specific antibodies.

Immunoprecipitation with HA. NIH/3T3 cells stably expressing TaPIN1 or TBL3 cells transiently expressing the FBW7 constructs were lysed in the following buffer: 20 mM Tris HCl pH 8, 150 mM NaCl, 0.6% NP-40 and 2 mM EDTA. Protein complexes were affinity-purified on anti-HA antibody-conjugated agarose

(Sigma, ref: A2095) for NIH/3T3 lysates or on anti-Flag antibody-conjugated agarose (Sigma, ref: A2220) for bovine lysates and eluted with the HA peptide or Flag peptide, respectively. After five washes, immunopurified complexes were resolved on 4–12% SDS–PAGE Bis-Tris acrylamide gradient gel in MOPS buffer (Invitrogen, ref: NP 0322 BOX, NP0001-02, respectively).

Immunoprecipitation with ubiquitin. Cells were treated for 3 h at 37 °C with 20 μM MG132 and lysed for 10 min on ice in the following buffer: 150 mM NaCl, 1% Nonidet P-40, 0.5% deoxycholate, 0.1% SDS, 50 mM Tris HCl pH 7.5, 20 mM NEM, 5 mM iodoacetamide, 100 μM MG132, 2 mg ml⁻¹ Pefabloc SC (Roche) and 5 μg ml⁻¹ each aprotinin, leupeptin, pepstatin. Equal amounts of total cellular proteins were immunoprecipitated with rabbit anti-c-JUN (E254) (Abcam, ref: ab32137) or rabbit anti-FBW7 (Bethyl Laboratories, ref: A301-721A), coupled to protein G sepharose beads (Sigma, ref: P3296) for 90 min at 4 °C. After three washes, immunoprecipitated proteins were eluted in Laemmli sample buffer at 95 °C for 5 min, resolved by SDS–PAGE and analysed by western blot using the indicated antibodies. Immunoprecipitation was repeated for three independent biological replicates.

Cycloheximide chase assay. Infected bovine cells (TBL3) were treated for 72 h with buparvaquone, juglone or DTM and transiently transfected with the indicated siRNA. Then, cells were treated for 30, 60 or 120 min with 100 mg ml⁻¹ cycloheximide. Cells were lysed in Laemmli sample buffer, resolved by SDS–PAGE and analysed by western blot using the indicated antibodies. Relative quantification indicates the c-JUN/tubulin ratios calculated with Image J software (NIH) and c-JUN levels at time 0 were set as 1. Cycloheximide chase experiments were repeated for four independent biological replicates.

Viability assays. 1×10^4 cells were plated in 96-well plates in triplicate and buparvaquone, juglone or DTM was added. Cell viability was measured after 72 h using the Cell proliferation Kit II–XTT (Roche) and the GloMax-Multi Detection System (Promega).

Data and statistical analysis. The GraphPad PRISM 6 program (GraphPad Software) was used for statistics. The results presented in all the figures represent the average ± s.d. of at least three independent experiments. Statistical analysis was performed using the one-way analysis of variance (ANOVA) and multiple comparisons test. P values were corrected for multiple comparisons using the Bonferroni correction based on the total overall number of pairwise comparisons for Fig. 1b. P values were calculated using the approach of Dunnett for multiple comparisons with the TaPIN1 wild type for Fig. 2c. For Figs 2b, d–f and 3c–e, P values were corrected using the Dunnett multiple comparisons with the control. P values with the Bonferroni method based on the number of pairwise comparisons were calculated for Fig. 4e. The statistics in Fig. 4h used the Dunnett procedure. Finally, an unpaired Mann–Whitney test was performed for the zebrafish experiments to analyse the significant difference between the control and treatment groups. The SPSS 19.0 program (SPSS) was used for statistics in Extended Data Figs 1–10. The results presented in Extended Data Figs 1–10 represent the average ± s.d. of at least three independent experiments. P values of <0.05 were considered statistically significant.

Bioinformatic screen. On 10 September 2011, search strategies at EuPathDB[32] were used to search for all T. annulata protein-encoding genes with predicted signal peptides (SignalP 2.0). Six-hundred and eighty-nine genes were returned. One-hundred and thirty-eight of these were found to have a predicted signal peptide only in T. annulata and not in their T. gondii orthologues. Among these proteins, we excluded (1) hypothetical proteins, (2) proteins that are not expressed at the macroschizont stage, and (3) proteins that are predicted to be targeted to the apicoplast of the parasite. We obtained 33 proteins, as shown in Extended Data Fig. 1a. This search strategy was repeated on 22 April 2013 to ensure that results were consistent with any EuPathDB updates. All 33 proteins from Extended Data Fig. 1a were returned in the updated search.

Analysis of protein structures and docking computations. Homology models of TaPIN1 wild type and the A53P mutant were built with the online server EsyPred[33]. The experimental structure of hPIN1 (ref. 34) co-crystallized with a dipeptide Ala-Pro (resolution 1.35 Å) was used as a template. To check the protonation state of the protein titratable groups, we used our online server PCE[35]. The most likely binding pocket areas for TaPIN1 wild type were predicted with FTMap and investigation of side-chain flexibility (if any) in the area of the predicted binding cavities was carried out with the server FTFlex[36]. The two-dimensional structures of juglone, buparvaquone and DTM were obtained from PubChem and the three-dimensional structures were generated with our package DG-AMMOS[37]. Three docking tools, Surflex[38], Molegro Virtual Docker[39] and our tool MS-DOCK[40], were used to search for possible poses of these compounds in hPIN1, for TaPIN1 wild type and the A53P mutant. Calibration of our docking tools was performed on the thiol-stress sensing regulator co-crystallized with a quinone molecule, which in this structure is covalently attached to a Cys residue (Protein Data Bank accession 4HQM)[41]. Visualization was carried out with PyMol and figures were also prepared with this molecular viewer package.

LETTER

β–Lactam formation by a non–ribosomal peptide synthetase during antibiotic biosynthesis

Nicole M. Gaudelli[1]†, Darcie H. Long[1] & Craig A. Townsend[1]

Non-ribosomal peptide synthetases are giant enzymes composed of modules that house repeated sets of functional domains, which select, activate and couple amino acids drawn from a pool of nearly 500 potential building blocks[1]. The structurally and stereochemically diverse peptides generated in this manner underlie the biosynthesis of a large sector of natural products. Many of their derived metabolites are bioactive such as the antibiotics vancomycin, bacitracin, daptomycin and the β-lactam-containing penicillins, cephalosporins and nocardicins. Penicillins and cephalosporins are synthesized from a classically derived non-ribosomal peptide synthetase tripeptide (from δ-(L-α-aminoadipyl)–L-cysteinyl–D-valine synthetase)[2]. Here we report an unprecedented non-ribosomal peptide synthetase activity that both assembles a serine-containing peptide and mediates its cyclization to the critical β-lactam ring of the nocardicin family of antibiotics. A histidine-rich condensation domain, which typically performs peptide bond formation during product assembly, also synthesizes the embedded four-membered ring. We propose a mechanism, and describe supporting experiments, that is distinct from the pathways that have evolved to the three other β-lactam antibiotic families: penicillin/cephalosporins, clavams and carbapenems. These findings raise the possibility that β-lactam rings can be regio- and stereospecifically integrated into engineered peptides for application as, for example, targeted protease inactivators[3,4].

Despite their widespread use for more than half a century, the β-lactam antibiotics, represented most familiarly by the semi-synthetic penicillins and cephalosporins, remain the most frequently prescribed anti-infectives in human medicine[5,6]. Four structurally distinct clans occur naturally, and the more recently discovered of these and their synthetic variants are of increasing importance to combat the rising spectre of antibiotic-resistant infectious diseases[7,8]. Members of this group of antibiotics contain monocyclic and fused bicyclic β-lactams whose high energy, strained-ring skeletons are essential to their antimicrobial activities. Markedly different but chemically efficient biosynthetic pathways have evolved to each of the penicillin and cephalosporin (for example isopenicillin N and cephalosporin C)[9], clavulanic acid[10] and carbapenem (for example thienamycin)[11] groups (Fig. 1a, b). Ironically, the fourth and structurally simplest clan of monocyclic β-lactams, exemplified by nocardicin G (Fig. 2b), has long remained an unsolved problem[12,13].

The nocardicin non-ribosomal peptide synthetase (NRPS) encompasses two megaenzymes, NocA and NocB, which together comprise five modules (Fig. 2a). Each module contains an adenylation (A) domain that binds ATP, selects its cognate building block and performs substrate acyl adenylation. The activated amino acid is then translocated as its aminoacyl thioester to the 4′-phosphopantetheine 'arm' of the downstream post-translationally modified peptidyl carrier protein (PCP). Condensation (C) domains mediate substrate inter-module amide bond formation to yield peptides of length and sequence defined by the NRPS(s). An epimerization (E) domain is embedded in module 3, which converts its associated amino-acid residue from the L- to the D- configuration. Finally, catalytic turnover of the NRPS is achieved

through disconnection of the final peptide product by the carboxy-terminal (C-terminal) thioesterase (TE) domain.

Although the roles of NocA and NocB in nocardicin A biosynthesis have remained enigmatic, it was established early on that the O-homoseryl terminus of nocardicin A is derived in an unusual transfer reaction from S-adenosyl-L-methionine[14] and that the β-lactam carbons arise from L-serine (Fig. 2b)[15]. The two modified p-(hydroxyphenyl)glycine (pHPG) units originate as catabolic products of L-tyrosine[16]. Nocardicin G (Fig. 2b), the simplest of the nocardicins, is a key pathway intermediate to nocardicin A[17]. As a consequence, it was initially thought that modules 1 and 2 of NocA were inactive to account for an apparent tripeptide NRPS precursor to nocardicin G. Subsequent experimentation, however, demonstrated that all five modules of NocA and NocB are essential to antibiotic production[12], and careful analysis of each dissected A domain gave the predicted product as L-pHPG–L-Arg–D-pHPG–L-Ser–L-pHPG, a pentapeptide[13]. The role of the amino-terminal (N-terminal) L-pHPG–L-Arg in the biosynthesis was unclear as was the means by which the C-terminal pHPG epimerized from the L- to the D-configuration present in nocardicin G. Recent experiments with the nocardicin thioesterase domain (NocTE) shed light on these questions and defined the central problem of β-lactam formation. A series of predicted
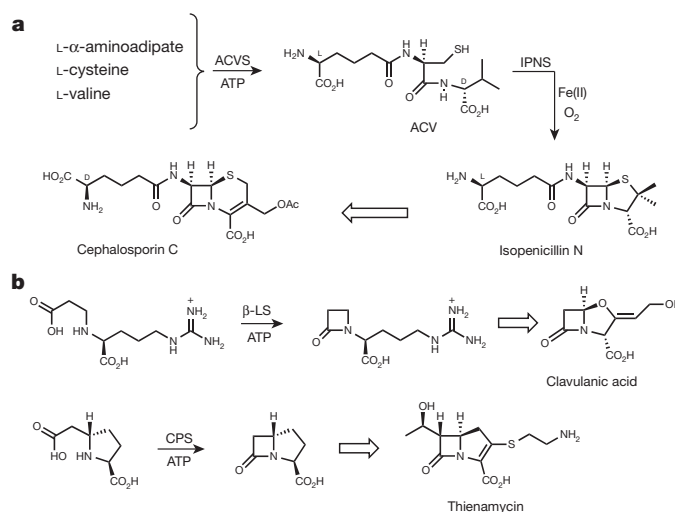
Figure 1 | Representative members of the family of β-lactam antibiotics. a, ACV (δ-(L-α-aminoadipic acid)–L-cysteine–D-valine) is an NRPS-derived tripeptide from ACV synthetase (ACVS). Isopenicillin N synthase (IPNS) catalyses oxidative β-lactam formation and bicyclization of ACV to form isopenicillin N with a single molecule of dioxygen and release of two molecules of water. Cephalosporin C is derived after isopenicillin N is epimerized to penicillin N and oxidative ring expansion occurs. **b**, The clavams and carbapenems are exemplified by clavulanic acid and thienamycin, respectively. Formation of the β-lactam ring that ultimately appears in clavulanic acid and thienamycin is catalysed by β-lactam synthetase (β-LS) and carbapenam synthetase (CPS), respectively, where transiently formed acyl adenylates are cyclized to β-lactam containing pathway intermediates, AMP and inorganic diphosphate.

[1]Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, USA.
†Present address: Department of Chemistry & Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA.
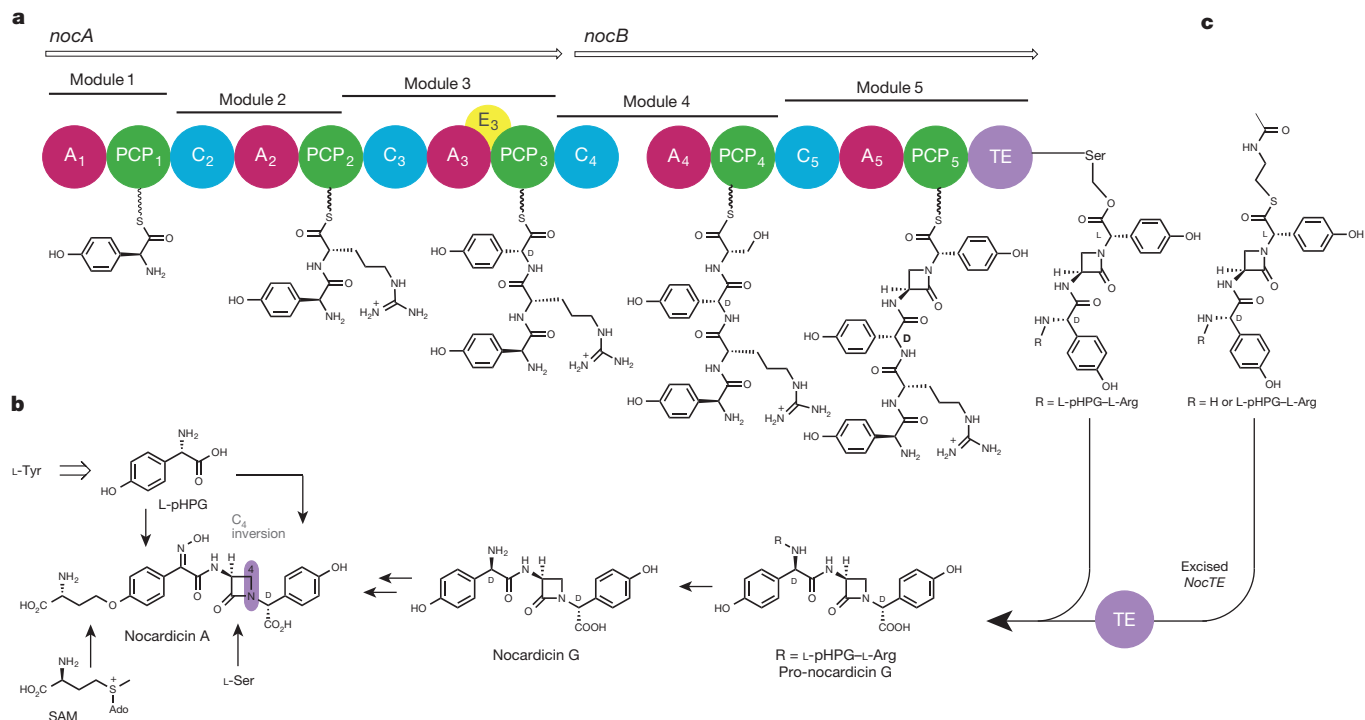
Figure 2 | Biosynthesis of nocardicin A. a, The nocardicin NRPS contains five modules, encoded by *nocA* and *nocB*, which together activate and condense in order L-pHPG, L-Arg, D-pHPG, L-Ser and L-pHPG. The thioesterase domain catalyses C-terminal pHPG epimerization and hydrolysis yielding pro-nocardicin G. **b**, Whole-cell experiments showed nocardicin A is derived from two units of L-pHPG, L-Ser and *S*-adenosyl-L-methionine, and from nocardicin G. **c**, Previous experiments with excised thioesterase demonstrate that tri- and pentapeptide thioesters containing a preformed β-lactam ring are converted to nocardicin G (R = H) and pro-nocardicin G (R = L-pHPG–L-Arg), respectively.

tri- and pentapeptide and potential seryl *O*-activated peptide thioesters all failed to undergo hydrolysis at rates greater than controls. On the other hand, the corresponding tri- and pentapeptide thioesters now bearing a preformed β-lactam ring from cyclization of the seryl residue were not only rapidly hydrolysed but also completely epimerized to the C-terminal D-stereochemistry (Fig. 2c)[18]. NRPS epimerase activity by a TE domain was unprecedented, but this specific instance is due to to the anomalously high acidity of a pHPG α-hydrogen relative to other α-amino acids[18]. Competition experiments established that the L,L,D,L,L-pentapeptide β-lactam thioester is the preferred NocTE substrate[18], a finding fully in accord with the requirement that all five modules of NocA/B are necessary for nocardicin biosynthesis.

Although NocTE catalyses C-terminal epimerization and hydrolytic product release, it was not observed to mediate β-lactam synthesis[18]. Azetidinone formation, therefore, must logically occur upstream on the NRPS after introduction of the last pHPG unit in module 5 from which the β-lactam ring nitrogen arises. In principle, formation of the embedded β-lactam ring could take place either in *cis* in this module or occur in *trans*. The latter alternative invokes the action of auxiliary enzyme(s), which are increasingly precedented in NRPS biochemistry[19]. Among the mechanisms that can be visualized are in *trans* activation of the seryl hydroxyl group by, for example, phosphorylation or acylation, and intramolecular nucleophilic substitution (S_Ni) by the adjacent amide to form the critical C4–N bond, a process well supported by chemical precedent[20] and consistent with the observation of stereochemical inversion at the seryl β-carbon[16]. Bioinformatic analysis and biochemical experiments, however, did not point to candidate auxiliary enzyme(s) encoded by the nocardicin biosynthetic gene cluster[21]. As a consequence, experiments were first undertaken to probe the in *cis* strategy with unexpected results.

The termination module of NocB, module 5, is composed of four domains: C_5, A_5, PCP_5 and TE. This 144 kilodalton protein was heterologously expressed in *Escherichia coli* with a His6 tag and purified by affinity chromatography. Complete conversion to its corresponding *holo*

form was ensured by Sfp-mediated 4′-phosphopantetheinyl transfer from coenzyme A (CoASH)[22,23]. The final chemical transformations catalysed by the termination module were successfully reconstituted *in vitro* through incubation of the predicted tetrapeptide-modified PCP domain from module 4 (PCP_4) with *holo*-module 5. Bearing in mind that all five modules of NocA/B are required for production of nocardicin A in *Nocardia uniformis*, and that the β-lactam-containing pentapeptide is preferentially processed by NocTE over the corresponding tripeptide[18], L-pHPG–L-Arg–D-pHPG–L-Ser-CoA (**1**, Fig. 3a) was prepared (Supplementary Information) and linked to *apo*-PCP_4 in an Sfp-mediated transfer to create L-pHPG–L-Arg–D-pHPG–L-Ser-*S*-PCP_4 (**2**, Fig. 3b and Extended Data Fig. 1). It was anticipated that module 5 would activate L-pHPG in the presence of ATP and present this amino acid on PCP_5 for reaction with the tetrapeptide delivered to module 5 by PCP_4. Indeed, when *holo*-module 5, 10 equivalents of tetrapeptidyl-*S*-PCP_4 (**2**), L-pHPG and ATP were combined, smooth conversion to the pentapeptide β-lactam (pro-nocardicin G) was observed (Fig. 3c and Extended Data Fig. 2). Monitoring product formation by high-performance liquid chromatography (HPLC) in a time-course experiment and simultaneous consumption of tetrapeptidyl-*S*-PCP_4 (**2**) by electrospray ionization mass spectrometry (ESI–MS) revealed a 1:1 correlation in accord with full catalytic turnover (Fig. 3d). Control experiments lacking L-pHPG or L-pHPG and ATP showed no product formation.

In a negative control experiment, the *in vitro* reconstitution experiment was repeated with a point mutant of C_5 where the second histidine residue of the conserved active site HHxxxDG sequence, known to be essential for amide bond formation[24], was replaced by alanine (H792A). No new products were detected (Fig. 3c). To further define acceptable substrates for C_5, the L-pHPG–L-Arg–D-pHPG–L-Ser-*S*-pantetheine (**3**, Fig. 3a) substrate mimic was prepared (Supplementary Information) but did not yield pro-nocardicin G when incubated with *holo*-module 5, ATP and L-pHPG (Extended Data Fig. 3a, b), a result that emphasizes the critical importance PCP_4•C_5 domain•domain interaction plays to β-lactam formation. Next the dipeptide D-pHPG–L-Ser-CoA (**4**, Fig. 3a)
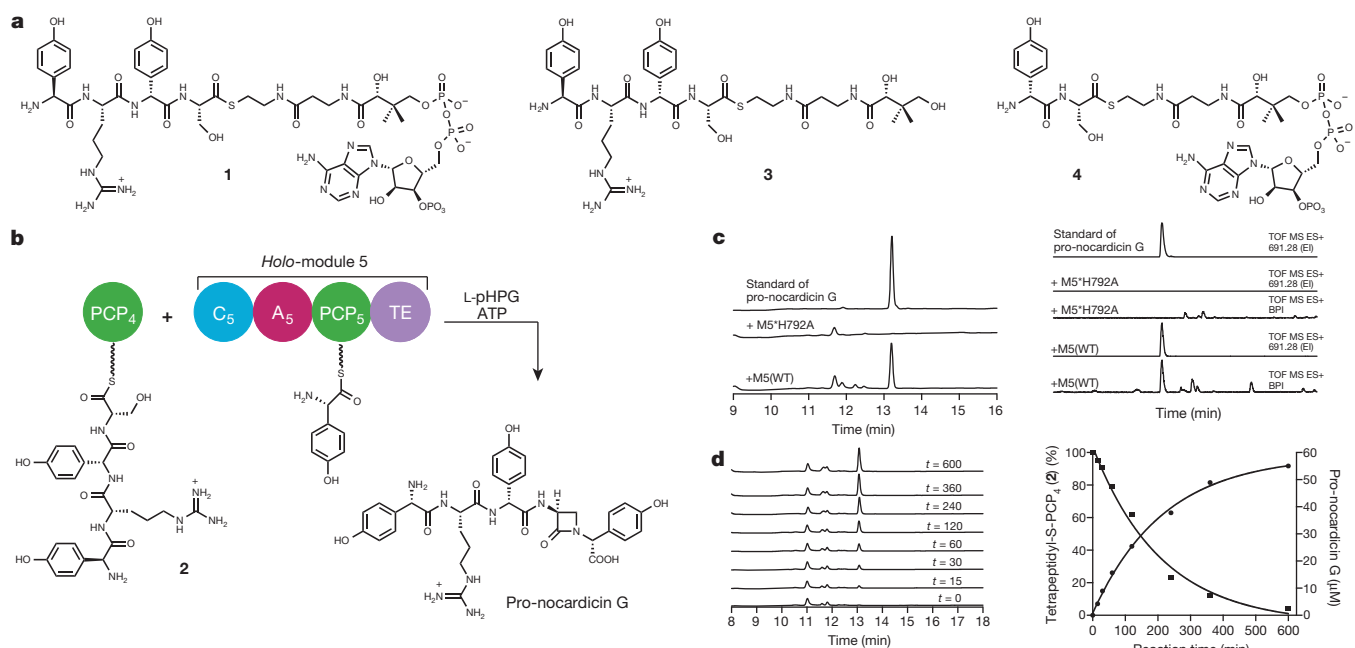
Figure 3 | Analysis of the reactions catalysed by module 5. a, Substrates used in this study. **b**, Incubation of tetrapeptidyl-*S*-PCP$_4$ **2** with *holo*-module 5, ATP and L-pHPG produced β-lactam containing pro-nocardicin G. **c**, Left: HPLC traces of products obtained after incubation of tetrapeptidyl-*S*-PCP$_4$ **2** and indicated construct, ATP and L-pHPG. This experiment was reproduced more than five times, and at least in duplicate for all other incubations. Pro-nocardicin G was observed in the wild-type reaction (+M5(WT)) but not in the mutant (+M5*H792A), verified by comparison with synthetic standard (top trace). Right: liquid chromatography–mass spectrometry (LC–MS) traces of products obtained after incubation of tetrapeptidyl-*S*-PCP$_4$ **2** and *holo*-module

5, ATP and L-pHPG. Pro-nocardicin G was observed in the wild-type reaction (+M5(WT)) but not in the mutant (+M5*H792A), verified by comparison with synthetic standard (top trace). BPI, base peak ion; TOF, time of flight; ES, electrospray ionization; EI, extracted ion. Calculated exact mass of pro-nocardicin G = 691.2835 [M + H]$^+$. **d**, Left: time-course study of tetrapeptidyl-*S*-PCP$_4$ **2** and *holo*-module 5 supplemented with L-pHPG and ATP. Appearance of pro-nocardicin G was analysed by HPLC. Right: plots of pro-nocardicin G production (HPLC) and the corresponding conversion of **2** to the unloaded *holo*-PCP$_4$ (ESI–MS). Less than 10% hydrolysis of **2** was observed during the 10-h experiment.

was prepared (Supplementary Information) and loaded onto *apo*-PCP$_4$ as before to afford D-pHPG–L-Ser-*S*-PCP$_4$ (**5**, Extended Data Fig. 4). When this construct was generated in the presence of *holo*-module 5, L-pHPG and ATP, nocardicin G was not detected (Extended Data Fig. 3c, d). These data suggest that the L-pHPG–L-Arg 'leader' present in the tetrapeptidyl-*S*-PCP$_4$ **2** plays a vital role in the binding and/or recognition of the upstream tetrapeptidyl intermediate in C$_5$, enabling peptide extension and β-lactam formation to occur.

Examination of the primary sequence of C$_5$ showed no unusual insertions or deletions except that, in addition to the conserved HHxxxDG catalytic motif emblematic of condensation domains, a third His residue (H790) lies directly upstream of the His dyad (Extended Data Fig. 5). Sequence analysis also revealed features of a ^{D}C$_L$ domain despite receiving an L-seryl tetrapeptide from PCP$_4$ (Extended Data Table 1). We propose a mechanism in which His 790 catalyses β-elimination of hydroxide (water) from the seryl residue of the PCP$_4$-bound tetrapeptidyl-thioester and PCP$_5$-tethered L-pHPG achieves β-addition with overall inversion of configuration at the seryl(dehydroalanyl) β-carbon dictated by earlier stereochemical experiments (Fig. 4a)[25]. The transient loss of the L-seryl stereocentre during the β-elimination/addition may account for the ^{D}C$_L$ characteristics of C$_5$. The resulting β-aminothioester **6** is then proposed to undergo unconventional amide bond cyclization (allowed 4-*exo*-trig), thermodynamically driven by amide bond formation from the active PCP$_4$ thioester. The PCP$_5$-bound pentapeptide β-lactam (pro-*epi*-nocardicin G) is poised for delivery to NocTE for C-terminal epimerization and hydrolytic product release.

To support this mechanistic hypothesis, we prepared a mutant of module 5 in which the tentative catalytic His was replaced by alanine (M5*H790A). Repeating the experiments with PCP$_4$-bound tetrapeptide **2** and mutant M5*H790A, L-pHPG and ATP yielded no product (Extended Data Fig. 6). Similarly site-specific mutation of the His residue typically involved in peptide bond formation (M5*H792A) also gave no reaction, as anticipated. In a further test of the proposed mechanism,

the reactive dehydroalanyl tetrapeptide intermediate (**6**, Fig. 4b) was synthesized (Supplementary Information) and used in an Sfp-catalysed reaction to afford the corresponding L-pHPG–L-Arg–D-pHPG–dehydroalanyl-*S*-PCP$_4$ substrate (**7**, Fig. 4c and Extended Data Fig. 7). In the course of preparing this sensitive material, it was discovered that the addition of sulphur, phosphorus and nitrogen nucleophiles occurred preferentially 1,4 rather than 1,2 in keeping with the hypothetical reactivity posed in Fig. 4a. When PCP$_4$-bound dehydroalanyl tetrapeptide **7** was incubated with wild-type *holo*-module 5, L-pHPG and ATP, β-lactam formation was once again observed (Fig. 4d and Extended Data Figs 8 and 9). Further insight into this process was afforded by the M5*H790A mutant, which did not support complete reaction of the dehydroalanyl substrate to the β-lactam product (Fig. 4d). The proposed catalytic residue H790 must not only act as a base to promote β-elimination but also serve as the acid to consummate amine (L-pHPG-*S*-PCP$_5$) β-addition. Although interfering with the proper cycling of the protonation state of the enzyme can be partly compensated in the wild-type protein, it cannot in the M5*H790A mutant.

NRPS C domains are pseudodimeric proteins whose N- and C-terminal subdomains are joined to form an extended V-shaped substrate channel that accommodates the donor and acceptor aminoacyl reactants, each delivered by extended pantetheinyl 'arms' from proximal PCP domains[26]. The centrally located HHxxxDG motif promotes peptide bond formation and transfer of the growing peptide chain to the downstream PCP domain. The unprecedented β-lactam formation catalysed by C$_5$ is distinct from the iron-mediated oxidative cyclization to penicillin (Fig. 1a) and the ATP-driven β-amino-acid closures that lead ultimately to clavulanic acid and all of the carbapenems (Fig. 1b)[27], and it does not correlate to the heterocyclization of serine residues to oxazolidine rings[28]. There is no stereoelectronic imperative that β-elimination/addition reactions must occur with overall retention of stereochemistry[29]. It can be readily appreciated that departure of the seryl OH and conjugate addition of the pHPG amine can take place on opposite faces of the
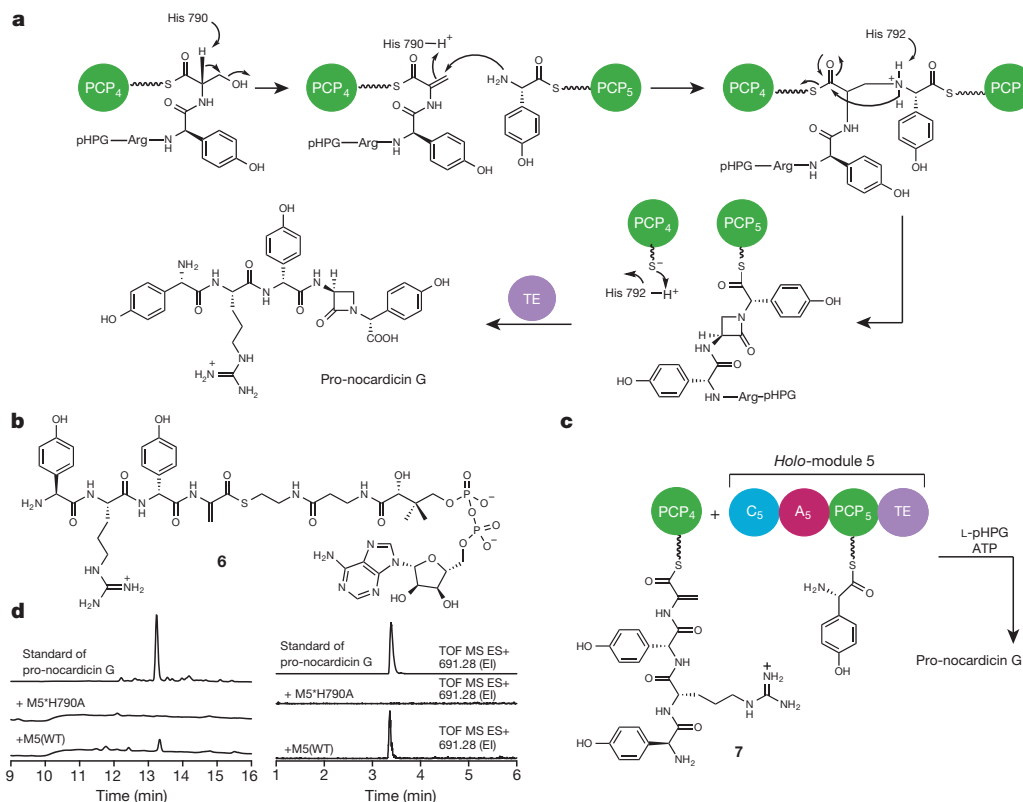
Figure 4 | Proposed β-lactam formation mechanism. a, Proposed mechanism of β-lactam formation in C$_5$. Tentative catalytic roles of histidine residues are indicated. **b,** Substrate used in this study. **c,** Incubation of dehydroalanyl tetrapeptidyl-S-PCP$_4$ **7** with holo-module 5, ATP and L-pHPG gave pro-nocardicin G. **d,** Left: HPLC traces of products obtained after incubation of dehydroalanyl tetrapeptidyl-S-PCP$_4$ **7** and indicated holo-module 5, ATP and L-pHPG. Pro-nocardicin G was observed in the wild-type reaction

(+M5(WT)) but not in the mutant (+M5*H790A), verified by comparison with synthetic standard (top trace). Right: LC–MS traces of products obtained after incubation of dehydroalanyl tetrapeptidyl-S-PCP$_4$ **7** and indicated holo-module 5 construct, ATP and L-pHPG. Pro-nocardicin G was observed in the wild-type reaction (+M5(WT)) but not in the mutant (+M5*H790A), verified by comparison with synthetic standard (top trace).

dehydroalanyl intermediate **7** to achieve configurational inversion. This facile β-elimination from a thioester-bound seryl residue stands in contrast to the generation of dehydroalanyl components and lanthionine bridges in ribosomally synthesized and post-translationally modified natural products (RiPPs)[30] where a prior O-phosphorylation or glutamylation intervenes to facilitate this elimination step. Impressive synthetic efficiency is achieved by the NocB termination module where previously unknown NRPS catalytic capabilities are captured in a non-oxidative route to β-lactams. Parsing a universe of more than 25,000 NRPS sequences to those that contain both a C domain bearing a HHHxxxDG motif and an immediately upstream A domain that is confidently predicted to activate Ser (or Thr) yielded only four Ser hits, two of which are known nocardicin producers[13], and, interestingly, five Thr hits (Extended Data Table 1). The products of the last seven, and whether they are β-lactam containing or not, are not known. The exceeding rarity of even potential β-lactam synthesis by an NRPS is emphasized by these findings, but its discovery at once expands the engineering and synthesis goals that can now be contemplated for this versatile class of giant modular enzymes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Walsh, C. T., O'Brien, R. V. & Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew. Chem. Int. Ed.* **52,** 7098–7124 (2013).

2. Banko, G., Demain, A. L. & Wolfe, S. δ-(L-α-Aminoadipyl)-L-cysteinyl-D-valine synthetase (ACV synthetase): a multifunctional enzyme with broad substrate specificity for the synthesis of penicillin and cephalosporin precursors. *J. Am. Chem. Soc.* **109,** 2858–2860 (1987).

3. Buller, A. R. & Townsend, C. A. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc. Natl Acad. Sci. USA* **110,** E653–E661 (2013).

4. Galletti, P. & Giacomini, D. Monocyclic β-lactams: new structures for new biological activities. *Curr. Med. Chem.* **18,** 4265–4283 (2011).

5. Hamad, B. The antibiotics market. *Nature Rev. Drug Discov.* **9,** 675–676 (2010).

6. Demain, A. L. Antibiotics: natural products essential to human health. *Med. Res. Rev.* **29,** 821–842 (2009).

7. McKenna, M. Antibiotic resistance: the last resort. *Nature* **499,** 394–396 (2013).

8. Livermore, D. M. Has the era of untreatable infections arrived? *J. Antimicrob. Chemother.* **64,** i29–i36 (2009).

9. Burzlaff, N. I. et al. The reaction cycle of isopenicillin N synthase observed by X-ray diffraction. *Nature* **401,** 721–724 (1999).

10. Bachmann, B. O., Li, R. & Townsend, C. A. β-Lactam synthetase: a new biosynthetic enzyme. *Proc. Natl Acad. Sci. USA* **95,** 9082–9086 (1998).

11. Freeman, M. F., Moshos, K. A., Bodner, M. J., Li, R. & Townsend, C. A. Four enzymes define the incorporation of coenzyme A in thienamycin biosynthesis. *Proc. Natl Acad. Sci. USA* **105,** 11128–11133 (2008).

12. Davidsen, J. M. & Townsend, C. A. In vivo characterization of nonribosomal peptide synthetases NocA and NocB in the biosynthesis of nocardicin A. *Chem. Biol.* **19,** 297–306 (2012).

13. Davidsen, J. M., Bartley, D. M. & Townsend, C. A. Non-ribosomal propeptide precursor in nocardicin A biosynthesis predicted from adenylation domain specificity dependent on the MbtH family protein NocI. *J. Am. Chem. Soc.* **135,** 1749–1759 (2013).

14. Reeve, A. M., Breazeale, S. D. & Townsend, C. A. Purification, characterization, and cloning of an S-adenosylmethionine-dependent 3-amino-3-carboxypropyltransferase in nocardicin biosynthesis. *J. Biol. Chem.* **273,** 30695–30703 (1998).

15. Townsend, C. A. & Brown, A. M. Nocardicin A: biosynthetic experiments with amino acid precursors. *J. Am. Chem. Soc.* **105,** 913–918 (1983).

16. Hubbard, B. K., Thomas, M. G. & Walsh, C. T. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics. *Chem. Biol.* **7,** 931–942 (2000).

17. Townsend, C. A. & Wilson, B. A. The role of nocardicin G in nocardicin A biosynthesis. *J. Am. Chem. Soc.* **110,** 3320–3321 (1988).
18. Gaudelli, N. M. & Townsend, C. A. Epimerization and substrate gating by a TE domain in β-lactam antibiotic biosynthesis. *Nature Chem. Biol.* **10,** 251–258 (2014).
19. Walsh, C. T. *et al.* Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Curr. Opin. Struct. Biol.* **5,** 525–534 (2001).
20. Salituro, G. M. & Townsend, C. A. Total syntheses of (−)-nocardicins A–G: a biogenetic approach. *J. Am. Chem. Soc.* **112,** 760–770 (1990).
21. Gunsior, M. *et al.* The biosynthetic gene cluster for a monocyclic β-lactam antibiotic, nocardicin A. *Chem. Biol.* **11,** 927–938 (2004).
22. Belshaw, P. J., Walsh, C. T. & Stachelhaus, T. Aminoacyl-CoAs as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science* **284,** 486–489 (1999).
23. Quadri, L. E. N. *et al.* Characterization of Sfp, a *Bacillus subtilis* phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **37,** 1585–1595 (1998).
24. Roche, E. D. & Walsh, C. T. Dissection of the EntF condensation domain boundary and active site residues in nonribosomal peptide synthesis. *Biochemistry* **42,** 1334–1344 (2003).
25. Townsend, C. A., Brown, A. M. & Nguyen, L. T. Nocardicin A: stereochemical and biomimetic studies of monocyclic β-lactam formation. *J. Am. Chem. Soc.* **105,** 919–927 (1983).
26. Samel, S. A., Schoenafinger, G., Knappe, T. A., Marahiel, M. A. & Essen, L.-O. Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure* **15,** 781–792 (2007).
27. Tahlan, K. & Jensen, S. E. Origins of the β-lactam rings in natural products. *J. Antibiot. (Tokyo)* **66,** 401–410 (2013).
28. Sattely, E. S. & Walsh, C. T. A latent oxazoline electrophile for N-O-C bond formation in pseudomonine biosynthesis. *J. Am. Chem. Soc.* **130,** 12282–12284 (2008).
29. Silverman, R. B. *The Organic Chemistry of Enzyme-Catalyzed Reactions* revised edn, 399–452 (Academic, 2002).
30. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30,** 108–160 (2013).

METHODS

No statistical methods were used to predetermine sample size.

Synthesis of all compounds used in this study can be found in the associated Supplementary Information.

General methods. Analytical HPLC analyses of enzymatic reactions were performed on an Agilent model 1200 HPLC equipped with a multi-wavelength ultraviolet–visible detector in conjunction with a reverse-phase Phenomenex Luna 5u phenyl/hexyl analytical column (250 mm × 4.60 mm internal diameter). Water + ACN + 0.1% TFA: 0–5 min isocratic 93% water + 7% ACN + 0.1% TFA, 5–22 min gradient 7–50% ACN + 0.1% TFA, 22–25 min gradient 50–7% ACN + 0.1% TFA, 25–35 min isocratic 93% water + 7% ACN + 0.1% TFA. Flow rate = 1.0 ml min^{-1}.

Ultra-performance liquid chromatography (UPLC)–HRMS samples were analysed on a Waters Acquity H-Class UPLC system equipped with a multi-wavelength ultraviolet–visible diode array detector in conjunction with a Waters Acquity BEH UPLC column packed with an ethylene bridged hybrid C-18 stationary phase (2.1 mm × 50 mm, 1.7 μm) in tandem with HRMS analysis by a Waters Xevo-G2 Q-ToF ESI mass spectrometer. Mobile phase: 100% water + 0.1% formic acid 0–1 min, 1–7.5 min 80% ACN + 0.1% formic acid, 7.5–8.4 min isocratic 80% ACN + 0.1% formic acid, 8.4–10 min 100% water + 0.1% formic acid. Flow rate = 0.3 ml min^{-1}.

Cloning, expression and purification of His$_6$-module 5. The module 5 gene containing C-A-PCP$_5$-TE of the termination module in NocB was PCR amplified from the pMG0531 cosmid[21] containing *nocA* and *nocB* genes using the M5-forward and M5-reverse primers (Supplementary Table 1) and Herculase-HF DNA polymerase (Agilent Technologies). The resulting PCR product was incorporated into a pCRBlunt-TOPO subcloning vector (Invitrogen) and sequence verified (Johns Hopkins University Core Sequencing Facility). The pCRBlunt-M5 construct was digested with NdeI and HindIII (NEB) and ligated with T4 DNA ligase (NEB) into a similarly digested pET28b (Novagen) vector to create the corresponding N-terminal 6×-His fusion construct.

Apo-module 5 was expressed using the pET28b-M5 vector in Rosetta 2(DE3)/pLysS *E. coli* cells (Novagen) and cultured at 37 °C in 1 l of 2× YT broth supplemented with 50 μg ml^{-1} kanamycin and 50 μg ml^{-1} chloramphenicol. Upon reaching an absorbance at 600 nm of 0.7, the temperature of the culture was reduced to 4 °C for 1 h. The temperature of the culture was raised to 18 °C and expression was induced with 1 mM isopropyl α-D-thiogalactopyranoside (IPTG) and grown at 18 °C for 18 h.

The cells were harvested by centrifugation (5,000*g*, 15 min, 4 °C) and stored at −80 °C. Cells were thawed in lysis buffer (50 mM phosphate, 300 mM NaCl, pH 8.0) and disrupted by sonication (60% amplitude, 9 s on/off, 3 min) on ice. Cell debris was removed by centrifugation (25,000*g*, 30 min, 4 °C) and the clarified cell lysate was incubated with 2 ml of 50% suspension per litre of cell culture of TALON metal affinity resin (Clontech) for 1–2 h at 4 °C in a batch-binding format. The suspension was loaded onto a gravity column and washed with two volumes of lysis buffer. The desired protein was eluted with a stepwise gradient of imidazole (20–300 mM) in lysis buffer. Fractions containing the purified protein, as determined by SDS–PAGE with Coomassie staining, were pooled and dialysed against 3 l of assay buffer twice containing assay buffer (50 mM HEPES, 25 mM NaCl, pH 7.5). Protein concentrations were quantified by Bradford assay.

Construction, expression and purification of His$_6$-module 5 point mutants. Site-directed mutagenesis designed to alter His 790 (H790A) and His 792 (H792A) used the splicing by overlap extension method, from the pET28b-M5 vector, with the appropriate DNA primers for the desired mutant (Supplementary Table 2). The reverse primers used in these PCR reactions employed a native PstI restriction site in the *module 5* nucleotide sequence. The resulting extension-overlap PCR product was incorporated into a pCRBlunt-TOPO subcloning vector and sequence verified. The pCRBlunt-M5* mutant construct was digested with NdeI and PstI. This extension-overlap product was ligated into a similarly digested pET28b-M5 to provide the desired C domain mutant of full-length M5. Expression and purification

of mutant constructs was achieved similarly through procedures described for the wild-type protein.

Construction, expression and purification of His$_6$-PCP$_4$. Gene *pcp$_4$* from *nocB* was PCR amplified from the pMG0531 cosmid using the PCP$_4$-forward and PCP$_4$-reverse primers (Supplementary Table 1) and Herculase-HF DNA polymerase. The resulting PCR product was incorporated into a pCRBlunt-TOPO subcloning vector and sequence verified. The pCRBlunt-PCP$_4$ construct was digested with NdeI and NotI and ligated with T4 DNA ligase into a similarly digested pET28b vector to create the corresponding N-terminal 6×-His fusion construct. Expression and purification of the PCP$_4$ monodomain was achieved similarly through procedures described for the module 5 constructs.

***In vitro* reconstitution of module 5 activity.** Loading of peptidyl-*S*-CoA onto apo-PCP$_4$. The *apo*-PCP$_4$ constructs were converted to their *holo* forms by an Sfp-mediated transfer of the desired peptidyl-*S*-CoA substrate with the *apo*-PCP$_4$ construct. *Apo*-PCP$_4$ (200 μM) was incubated with 250 μM of desired peptidyl-*S*-CoA substrate in assay buffer supplemented with 10 mM MgCl$_2$. 4′-Phosphopantetheine transfer reactions were initiated by the addition of 2 μM of Sfp, and the enzymatic mixture was incubated for 45 min at room temperature. Excess peptidyl-*S*-CoA reagent was removed through serial dilutions of the reaction mixture. This was achieved by adding three volumes of assay buffer to the reaction mixture and concentrating the mixture back down to the initial volume using a 3k MWCO Amicon Ultra centrifugal filter (Millipore). This dilution procedure was repeated three times.

Generation of *holo*-module 5 construct. To a separate 1.5 ml tube, 20 μM of *apo*-module 5 construct (either wild type or mutant) was incubated with 40 μM coenzyme A in assay buffer supplemented with 10 mM MgCl$_2$. 4′-Phosphopantetheine transfer was initiated by the addition of 2 μM of Sfp and the reaction was left to stand for 45 min at room temperature. Excess CoA reagent was removed through serial dilutions of the enzymatic mixture as before.

Module-5-catalysed β-lactam formation. *Holo*-module 5 constructs were supplemented with 5 mM ATP and 2 mM L-pHPG and left to stand for 5 min in assay buffer. Condensation reactions were initiated by adding equal volumes of peptidyl-*S*-PCP$_4$ construct with *holo*-module 5 and left to stand for 2 h. The reaction contained 100 μM peptidyl-*S*-PCP$_4$, 10 μM *holo*-module 5 2.5 mM ATP and 1 mM L-pHPG in assay buffer. Proteins were removed by centrifugation through a 3k MWCO Amicon Ultra centrifugal filter. The filtrate was directly analysed by HPLC and products of interest were collected over multiple injections and concentrated by lyophilization. The concentrated samples were re-suspended in 70 μl of 95:5 water:ACN + 0.1% formic acid and directly analysed by LC–MS.

Incubation of tetrapeptidyl-*S*-pantetheine **3** with *holo*-module 5. Reactions in which pantetheinyl substrate **3** was substituted for the *holo*-PCP$_4$ construct contained 1 mM **3**, 10 μM *holo*-module 5 construct, 1 mM L-pHPG and 2.5 mM ATP and were left to stand for 2 h at room temperature in assay buffer. Reactions were quenched and analysed as described above.

31. Ehmann, D. E., Trauger, J. W., Stachelhaus, T. & Walsh, C. T. Aminoacyl-SNACs as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chem. Biol.* **7,** 765–772 (2000).
32. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** W362–W367 (2011).
33. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33,** 5799–5808 (2005).
34. Clugston, S. L., Sieber, S. A., Marahiel, M. A. & Walsh, C. T. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a (D)C(L) catalyst. *Biochemistry* **42,** 12095–12104 (2003).
35. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).

CORRECTIONS & AMENDMENTS

Corrigendum: Endocrinization of FGF1 produces a neomorphic and potent insulin sensitizer

Jae Myoung Suh, Johan W. Jonker, Maryam Ahmadian, Regina Goetz, Denise Lackey, Olivia Osborn, Zhifeng Huang, Weilin Liu, Eiji Yoshihara, Theo H. van Dijk, Rick Havinga, Weiwei Fan, Yun-Qiang Yin, Ruth T. Yu, Christopher Liddle, Annette R. Atkins, Jerrold M. Olefsky, Moosa Mohammadi, Michael Downes & Ronald M. Evans

This Letter should have declared the following competing financial interests: "The fibroblast growth factor (FGF) molecules and related methods of use reported in this study are covered in the following published patent applications and counterparts that derive priority: (1) PCT/US2011/032848, held by R.M.E., M.D., J.W.J., and J.M.S. (handled by Salk OTD); (2) PCT/US2013/044589, held by M.M., R.G., R.M.E., M.D. and J.M.S. (handled by NYU Office of Industrial Liaison/ Technology Transfer); (3) PCT/US2013/044594, held by M.M., R.G., R.M.E., M.D. and J.M.S. (handled by NYU Office of Industrial Liaison/ Technology Transfer); and (4) PCT/US2013/044592, held by M.M. and R.G. (handled by NYU Office of Industrial Liaison/Technology Transfer).".

CORRECTIONS & AMENDMENTS

Corrigendum: Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism

Fiona Cuskin, Elisabeth C. Lowe, Max J. Temple, Yanping Zhu, Elizabeth A. Cameron, Nicholas A. Pudlo, Nathan T. Porter, Karthik Urs, Andrew J. Thompson, Alan Cartmell, Artur Rogowski, Brian S. Hamilton, Rui Chen, Thomas J. Tolbert, Kathleen Piens, Debby Bracke, Wouter Vervecken, Zalihe Hakki, Gaetano Speciale, Jose L. Munõz–Munõz, Andrew Day, Maria J. Peña, Richard McLean, Michael D. Suits, Alisdair B. Boraston, Todd Atherly, Cherie J. Ziemer, Spencer J. Williams, Gideon J. Davies, D. Wade Abbott, Eric C. Martens & Harry J. Gilbert

In this Article focusing on the selfish metabolism of yeast mannan by Bacteroidetes, we also described a polysaccharide utilization locus (PUL) responsible for the degradation of high mannose mammalian *N*-glycan (HMNG) but omitted to cite two relevant papers[1,2], for which we apologise. Both studies describe a model for the degradation of complex biantennary *N*-glycans by Bacteroidetes in which the degradative enzymes are encoded by PULs. These studies[1,2] provide examples of how PULs can orchestrate *N*-glycan metabolism in addition to the HMNG PUL we describe in this Article. In all three papers it is proposed that *N*-glycan depolymerization occurs primarily in the periplasm.

1. Renzi, F. *et al.* The *N*-glycan glycoprotein deglycosylation complex (Gpd) from *Capnocytophaga canimorsus* deglycosylates human IgG. *PLoS Pathog.* **7,** e1002118 (2011).
2. Nihira, T. *et al.* Discovery of β-1,4-D-mannosyl-*N*-acetyl-D-glucosamine phosphorylase involved in the metabolism of *N*-glycans. *J. Biol. Chem.* **288,** 27366–27374 (2013).

CORRECTIONS & AMENDMENTS

Erratum: A new antibiotic kills pathogens without detectable resistance

Losee L. Ling, Tanja Schneider, Aaron J. Peoples,
Amy L. Spoering, Ina Engels, Brian P. Conlon, Anna Mueller,
Till F. Schäberle, Dallas E. Hughes, Slava Epstein, Michael Jones,
Linos Lazarides, Victoria A. Steadman, Douglas R. Cohen,
Cintia R. Felix, K. Ashley Fetterman, William P. Millett,
Anthony G. Nitti, Ashley M. Zullo, Chao Chen & Kim Lewis

In Fig. 3d of this Article, the '2:1' and '1:1' labels at the bottom of the panel were inadvertently switched during the production process; this figure has now been corrected in the online versions of the paper.

A MOST EXCEPTIONAL RESPONSE

Sometimes a drug causes a tumour to completely recede, but only in a tiny percentage of people. Scientists want to decipher such outlier responses for the benefit of all patients.

BY VIVIEN MARX

If Patient X were like most people with advanced bladder cancer, she would probably be dead by now. After her first diagnosis, she received standard chemotherapy. It failed. Then she entered a clinical trial for a drug that was originally approved to treat other tumour types: would it also work in metastatic bladder cancer? Apparently not — none of the other patients in the trial did well.

Yet Patient X thrived. Her tumour completely disappeared, says computational biologist Barry Taylor at Memorial Sloan Kettering Cancer Center (MSKCC) in New York, where Patient X was treated. Today, a little more than five years after treatment, she is healthy and has no evidence of disease[1].

Patient X (her identity is shielded to protect her privacy) is an exceptional responder, one of those rare individuals who have a dramatically positive response to a therapy that does little or nothing for most other patients. This response is not unique to cancer. Immunologists, for example, have discovered why some individuals can be HIV-positive and yet avoid the symptoms of AIDS.

By definition, exceptional responses are rare, which makes them hard to study. Their anecdotal nature seems to contradict the teachings on statistically sound results in biomedical research. In a clinical trial, even if there are several exceptional responders, a drug will ▶

▶ fail to achieve approval because it does not improve the health of the majority of patients. This means there has been little incentive for researchers or drug companies to investigate thoroughly why a few people respond so well.

But that neglect is starting to be addressed as more cases of exceptional responses in cancer reach the published scientific literature and techniques emerge for profiling patients at the molecular level[2]. In Patient X's case, genome sequencing revealed a mutation in her tumour that explains why her cancer is specifically vulnerable to the drug she received on the clinical trial[1]. Such successes indicate that searching for and profiling these patients can potentially help researchers to predict many other patients' responses to potential therapies.

The relatively new ability to comprehensively characterize a tumour's genome, transcriptome (its gene expression) and metabolome (its metabolic processes) increases the chance of discovering the reasons behind outlier results, says Kenneth Kinzler, a cancer researcher at the Johns Hopkins Kimmel Cancer Center in Baltimore, Maryland. "The hope is that a signal seen in an exceptional responder will be seen in other cancer patients and be a predictor of therapeutic response regardless of tumour type," he says.

THE EXCEPTIONAL PROFILE

There is no universally accepted definition of exceptional responders, says Barbara Conley of the US National Cancer Institute (NCI) in Rockville, Maryland. Conley directs the Exceptional Responders Initiative (ERI), which profiles these patients. The ERI considers a drug response to be exceptional when a tumour disappears or when a patient shows an exceptional response to treatment and lives longer than 90% of others treated similarly. In tough-to-treat and advanced cancers, an exceptional response is when treatment causes a tumour to regress by at least 30% for at least six months, but only in less than 10% of people on the same treatment.

In the case of Patient X, for example, her sequenced tumour genome revealed a mutation in a gene called tuberous sclerosis complex 1, which is one of several genes involved in a pathway that regulates cell growth and proliferation. The drug that worked for Patient X, but not for the other patients in the clinical trial, inhibits signalling in that pathway.

But that does not completely explain Patient X's exceptional response. Analysis of tumour samples from 13 other patients in her trial showed that four had a mutation in the same gene, but the drug gave them only a short reprieve. To get a better understanding of Patient X and other exceptional responders, the ERI wants to do comprehensive profiling of a wide variety of parameters, including

"The hope is that a signal seen in an exceptional responder will be seen in other cancer patients."

patients' clinical history, DNA changes, RNA levels of different genes (which reflect their activity) and metabolic pathways.

Taylor and his colleagues have long encountered the critique that studying exceptional responders is merely generalizing anecdotes. But even though published studies on exceptional responders are few, he says, "I think the weight of evidence has now shifted that view."

Vincent Miller, a former MSKCC oncologist, agrees that views about outliers are changing and thinks that many more such individuals might be found. Any oncologist has a handful of patients in whom cancer just melts away with no obvious explanation, says Miller, who is chief medical officer of Foundation Medicine in Cambridge, Massachusetts, a company that performs genomic analysis of samples from people with cancer. In January, the pharmaceutical company Roche, based in Basel, Switzerland, bought a majority stake in Foundation Medicine, which is also involved in the ERI.

The ERI encourages clinicians to get in touch if one of their patients has an exceptional reaction to a drug. At that point, a multidisciplinary review determines whether a more comprehensive profile is warranted, says Conley. In approved cases, and with the patient's consent, the physician sends in the complete medical record and a tumour sample. Around 160 submissions are currently under review. Conley and her team have been surprised to see submissions about established drugs as well

as drugs still under development.

The ERI makes sense only because large-scale sequencing efforts such as The Cancer Genome Atlas (TCGA) now offer huge data stores, says David Wheeler, who leads the ERI genome-analysis team at the human genome sequencing centre of Baylor College of Medicine in Houston, Texas. From Baylor, genome data will go to a database that is accessible by the research community.

The first few ERI samples are now beginning to arrive at Baylor, and researchers there are all set to potentially perform whole-genome sequencing using their newly arrived equipment — HiSeq X Ten Illumina sequencers. Whole-genome sequencing is ideal, says Wheeler, because it provides the most complete genomic information. But it also requires enough sample and plenty of time and money; so when the samples are smaller or when only ones with lower tumour purity are available, the team will just focus on protein-coding genes, which make up the exome.

For now, the ERI is in a pilot phase. If it proves successful, it could be scaled up by, for example, helping cancer treatment centres to forage for exceptional responders in their biobanks. But the pilot faces a few challenges.

One key issue is time, says Kristen Leraas, who is the sample coordinator at the biospecimen processing facility of the Nationwide Children's Hospital in Columbus, Ohio, where all of the ERI's samples are processed

Exceptional responders can help scientists to predict the responses of many other patients with cancer.

MEMORIAL SLOAN KETTERING CANCER CENTER

CLOCK-WATCHERS

The Nationwide Children's Hospital in Columbus, Ohio, deals with tumour samples from the Exceptional Responders Initiative of the US National Cancer Institute (NCI). Extracting DNA and RNA from the precious, tiny samples involves a race against the clock to avoid degradation of nucleic acid.

DAY 1

Processing of the newly arrived, tiny piece of tumour from an exceptional responder begins immediately

9:00 AM

Treatment centres across the United States submit cases to the NCI. After a review, tumour samples from patients who show a truly exceptional response are sent to Nationwide.
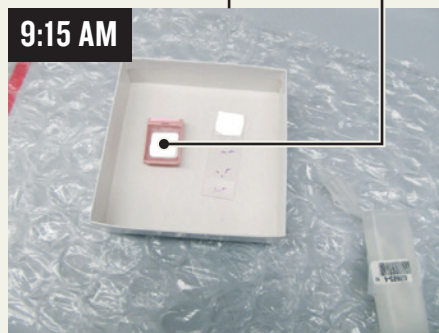
9:15 AM

The tumour tissue often arrives fixed in formalin and embedded in paraffin. This process can damage nucleic acids but allows for pathology review and long-term storage.
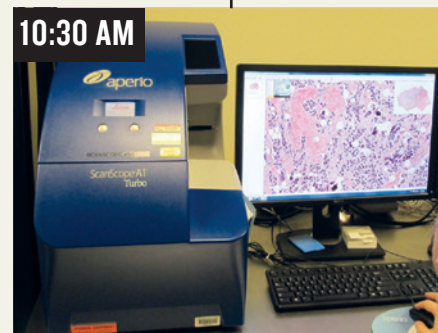
10:30 AM

In one of many quality-control and preparation steps during the first day, the tissue slide is scanned so a pathologist can confirm the tumour type, assess its quality and see the amount of tumour in the sample.

(see 'Clock-watchers'). When a sample comes in, she says, scientists have to race against the clock to process, standardize and prepare it for sequencing: DNA and RNA have to be extracted from the sample quickly to avoid any kind of degradation. "We pretend our hair is on fire and we make sure we extract right away."

Another challenge is that exceptional responses are unexpected, so the cancer centres sending tissue samples to Columbus do not collect them in a standardized way. One sample might be blood from someone with leukaemia, whereas another might come from a solid tumour. And unlike the case with the TCGA, it might arrive without a matched healthy tissue sample from the same person. A sample might be smaller than a pencil eraser and, in some cases — when it comes from a fine-needle biopsy, for example — it might even be invisible to the naked eye. TCGA samples weigh on average 260 milligrams, whereas "if we get 100 milligrams, that's a lot", says Jay Bowen, who directs logistics and data management at Nationwide's biospecimen processing facility. "Sometimes we make do with about 20 milligrams."

COMPREHENSIVE TESTING

The Nationwide laboratory's top priority with these samples is to extract enough nucleic acid to allow multiple analyses, including exome and messenger RNA sequencing. Some DNA is also sent to Foundation Medicine, where tests can detect and validate four classes of DNA alterations at once: substitutions of bases along the DNA strand, genetic insertions and deletions, changes in the number of copies of genes present in the genome and structural rearrangements[3]. Ideally, if the sample yields sufficient quantities of nucleic acid, whole-genome sequencing or

other types of tests, such as analysis of DNA methylation, can be performed.

A potential complication is that tumour samples taken during surgery or biopsy are often fixed in formalin and then embedded in paraffin. These formalin-fixed paraffin-embedded (FFPE) samples are standard in medical centres and are preferred by pathologists, who can easily shave off a thin slice when they want to study the tumour's cellular morphology under a microscope as part of diagnosis.

But this process can crosslink nucleic acids, and can also oxidize and shear these molecules, says molecular biologist Erik Zmuda, who directs molecular characterization tasks at the Nationwide's biospecimen processing facility. There is a risk that a genomic signal indicative of an exceptional response is actually an FFPE artefact. Thus, for studying the tumour's genome, researchers much prefer frozen tissue.

Zmuda and his colleagues at Nationwide and other institutions think they see a way to allow pathologists to continue to use their preferred FFPE preservation method while providing molecular biologists with the ability to profile a sample at the resolution they need. The team's idea is to find a telltale signature of FFPE artefacts in tumour samples, which would allow them to computationally mask these effects in the data. The team is developing an algorithm that would correct for the artefacts and thus make it easier to compare data from FFPE and frozen samples. That, in turn, could open up possibilities to retroactively analyse patient samples from pathology departments in any hospital. As well as helping the hunt for signals in outlier genomes, this method could also be adapted for use in genome analysis more generally when diagnosing and treating patients.

Other fields have a longer tradition than cancer research does of looking at exceptional responders, says Stephen Friend, a former director of the oncology division of pharmaceutical company Merck in Kenilworth, New Jersey. Early in the AIDS epidemic, for example, immunologists noticed that some people can be HIV-positive but lack symptoms. This exceptional biology was found to result from a mutation that changes a protein on the surface of the immune-cell type that HIV infects, thus stopping HIV from entering the cell[4].

Such links between a specific mutation and disease have sometimes led to targeted drugs. But genomics is not a black and white world in which certain mutations lead to the same clinical course in all patients, says Friend. Environmental factors and other genetic variants play their part too. This may be why these targeted drugs do not work in 100% of the patients with that mutation, he says.

Friend co-directs the Resilience Project (http://resilienceproject.me), which is geared towards finding outliers in many diseases[5]. The goal is to find people who harbour DNA changes that cause severe and rare childhood diseases, or that heighten cancer risk, but who have lived into healthy adulthood in spite of their genomes.

The programme is run by the non-profit organization Sage Bionetworks, which is based in Seattle, Washington, and is devoted to setting up platforms through which scientists can collaborate and share data. The Resilience Project currently consists of researchers from the Icahn School of Medicine at Mount Sinai Hospital in New York (conversations are also under way with the Gurdon Institute in Cambridge, UK). DNA analysis is in progress on samples from

The shaved slices roll up into scrolls

Genomic analysis will help researchers to understand what is special about exceptional responders

DAY 2

DAY 3

9:00 AM

The clock starts ticking in earnest when scientists shave off slices of the paraffin block that contains the tumour to begin extracting DNA and RNA. Exposure to air can alter molecules and even change genetic sequences.

10:00 AM

Nucleic acids must be isolated from the slices within 24 hours of the shaving to prevent possible damage caused by exposure to air.

12:00 PM

Nucleic acids are quality controlled and made ready for shipment. The extracted nucleic acids can be sent on for sequencing and analysis.

more than half-a-million donors, says Friend, who also directs Sage Bionetworks.

If the first analysis of the donor DNA reveals a mutation that could have killed the carrier, researchers can dig deeper into that person's genetics and biochemistry in an effort to understand their resilience. If one mutation is decisive, analysis can be quick, says Friend. But a mutation might act in conjunction with secondary mutations elsewhere in the genome. Searching for such mutation combinations is difficult, he says. But with an outlier genome in hand, researchers are at least trawling through a bucket of data, not an ocean of data.

Scientists tend to keep findings under wraps until they publish. But Friend thinks that analysis should be a collaborative task that is spread across multiple laboratories. This would increase the speed at which scientists can decipher which factors — be they genetic, immunological, environmental or a combination — have protected resilient individuals. "What I'm hoping is that we can get scientists to take it on as a sort of crowd-sourced federated approach," says Friend. "No one is paid to do that, no one owns the data."

RARE SIGNALS
In a clinical trial, scientists strive for numbers: making sure there are sufficient cases of disease and controls to see whether a drug is having an effect, for example. They look for global trends rather than focus on the outliers, says Gustavo Stolovitzky, a researcher for the technology firm IBM in Yorktown Heights, New York, who runs the Dialogue for Reverse Engineering Assessment and Methods, a research venture and competition that, for example, looks at how well different algorithms

predict the reaction of cancer cells to drugs[6].

By definition, outliers are too rare to have much statistical power, Stolovitzky says, and are usually dismissed as flukes. But conversely, he says, an exceptional response is a strong signal that is hard to miss. If many scientists hunt for exceptional responders in data from the ERI or the Resilience Project, perhaps 20 or even 50 cases can emerge. "That's starting to be something," says Stolovitzky. "It's a number we can do statistics with." If so, it may be possible to glimpse patterns that can help to explain how exceptional responders beat the odds.

In profiling outliers, scientists will not know which of the molecular signals is decisive, which is why comprehensive profiles are needed for everything — genome sequencing data, gene expression data, clinical data and other assay results. Comparing these profiles is tricky: for example, it can often be a challenge to compare genomic sequence, says Trey Ideker, a computational biologist at the University of California, San Diego. "We sequence this individual and they're a snowflake," he says — showing patterns that are unique even though the patients have the same type of cancer.

Ideker says that one approach to address that diversity is to view cancer as a disease of pathways, in which groups of genes act together to perform functions in the cell. When analysed on a pathway level, he says, patterns do emerge. For example, researchers may find that dissimilar-seeming mutations in a cancer all fall in a certain pathway, meaning that they all impair the same cell function.

These network patterns are not complete biochemical explanations of an exceptional response in cancer treatment, says Ideker, but they are indications of what to explore next.

Crucially, he says, by considering pathways, an exceptional responder becomes part of a group. Even if it is not a large group, the person is no longer an outlier.

Many patients could benefit from ventures to decipher the molecular profile of exceptional responders. A physician might realize that a drug that was not expected to do well in a given patient might actually be surprisingly suitable, says Taylor. This approach to cancer treatment complements an emerging idea that rather than focusing on the organ in which the tumour originated, treatments should be targeted to the molecular profile driving a given cancer.

For research on outliers to be of greatest help, the outlier cases must be rigorously selected. Only then can the analysis deliver sound results despite the fact that it remains a profile of only one person, says Friend. Taylor agrees, pointing out that molecular analysis of tumours from patients is increasingly possible and that there is growing acceptance of studying outlier patients. "Nevertheless," he says, "it requires that we stay focused on exploring the most significant outlier responses to ensure the greatest return for patients." ∎

Vivien Marx *is technology editor for* Nature *and* Nature Methods.

1. Iyer, G. *et al. Science* **338,** 221 (2012).
2. Wagle, N. *et al. N. Engl. J. Med.* **371,** 1426–1433 (2014).
3. Frampton, G. M. *et al. Nature Biotech.* **31,** 1023–1031 (2013).
4. Gorry, P. R. *et al. Lancet* **359,** 1832–1834 (2002).
5. Friend, S. H. & Schadt, E. E. *Science* **344,** 970–972 (2014).
6. Costello, J. C. *et al. Nature Biotech.* **32,** 1202–1212 (2015).

CAREERS

@NATUREJOBS Follow us on Twitter for the latest news and features

ARCTIC-IMAGES/GETTY

The Glaciological Society measures ice of glaciers each year for research now used to study global warming.

GLACIOLOGY

Climatology on thin ice

Ice-core scientists struggle to adapt as the subject of their research melts away.

BY NEIL SAVAGE

When Margit Schwikowski hiked up a glacier on the Svalbard islands in the Arctic Ocean a few years ago to collect ice samples for her climate research, she was gob-smacked. The Swiss analytical chemist had been to this site in 1997 to drill a core to test for trace gases and aerosols in her lab at the Paul Scherrer Institute in Villigen. But when she returned in 2009 for fresh samples, she could no longer reach the site: the warming glacier had cracked open and developed a yawning crevasse.

"You cannot go there," she says today. She has not tried to return since. It would be pointless, in any case — the crack had let in fresh snowfall and melt water, which then mixed with ice deeper in the glacier and confused any data that a scientist might try to extract.

Scientists who study the cryosphere — places on Earth that are sheathed in ice — are finding their jobs more difficult as the ice melts and glaciers recede. The subject that they study, climate change, poses intellectual challenges to their science and physical challenges to the way they approach fieldwork. Ice-core researchers need to be aware of how changing conditions affect the quality of their data. They must be prepared to go to more extreme environments to get their samples, and to face the challenges of those environments, such as dangerous terrain and low levels of oxygen.

And, unusually within fields of scientific study, ice-core scientists must find ways to ▶

▶ preserve the object of their study for future researchers before it vanishes forever — by stepping up efforts to collect and store the ice.

MEASURING THE MELT

Disappearing ice can complicate not just sample collection, but also analysis. Qianggong Zhang, an environmental geochemist at the Chinese Academy of Sciences in Beijing, knows this all too well. In 2005, he and colleagues at the academy's Institute of Tibetan Plateau Research climbed a glacier in Tibet to extract an ice core. Back at their lab, they characterized the ice layer by layer, measuring the concentrations of trapped gases and gathering other chemical information. But then they discovered that the top layers of ice were missing, rendering their work useless.

Scientists date ice cores by counting their layers, which vary as the seasons change and leave distinctive stratification. If the ice core is intact, the top layer should be the most-recent year; from there, researchers can tie what they learn to other information, such as records of temperature or precipitation. They can also look for signatures that serve as labels for specific years: atmospheric nuclear tests in the 1950s and 1960s, for instance, left a datable signature in glaciers worldwide, as did the 1986 Chernobyl meltdown in the then-Soviet Union.

Zhang's group had already done a lot of work before it noticed something amiss: none of its ice contained any trace of radioactive fallout. A core free of radiation indicated that the ice from those years had melted away — and Zhang's most recent layer of ice had to have come from a year earlier than 1950. But there was no good way to determine the top layer's age. "Our work on the other parameters is probably just a waste of time," Zhang laments. So the team's measurements sit in his computer, waiting for researchers to develop different dating methods. They wait, too, for a way to analyse layers that are chemically tainted by more-recent water. "We have to find new methods to dig out the information in the partially melted ice," Zhang says.

Some of that information could come, for example, from studying stable isotopes of oxygen that remain in partially melted ice, and remnants of insoluble material such as black carbon might help researchers to count layers. Although these clues might not be enough for scientists to obtain year-to-year data from the ice, they could enable them to retrieve average measurements from 5- or 10-year periods. It could also be possible for researchers to sift through the sediments of glacial lakes to retrieve material that has been washed out by ice melt, and treat that material as a proxy for the ice. But in samples that lack the radioactive signature, it

"We have to find new methods to dig out the information in the partially melted ice."

may never be possible to pinpoint specific years.

"Time no longer starts at the surface," says Lonnie Thompson, a palaeoclimatologist at the Byrd Polar Research Center at the Ohio State University in Columbus. He and his wife, Ellen Mosley-Thompson, have been collecting ice cores since the mid-1970s. He drilled an ice core from the Quelccaya ice cap in the Peruvian Andes in 1983, at which point no melting had occurred at altitudes above 5,000 metres. When he returned for another sample 20 years later, melting had altered the concentration of atmospheric isotopes in the top 40 metres of ice.

FREEZING FOR THE FUTURE

One way to get — and get to — a pure sample is to climb higher, where melting is not yet a problem. But that works only if there is actually somewhere to go. "In most cases, we can't go any higher. We can't get to a colder environment," says Douglas Hardy, a geoscientist at the University of Massachusetts Amherst.

Hardy places weather instruments on glaciers such as the one on Mount Kilimanjaro in Tanzania, which has shrunk by about 4 metres in the past 15 years. The instruments measure various meteorological conditions — temperature, humidity, precipitation rates and the amount of sunlight that strikes the glacier — and will help scientists to examine how those conditions affect growth or shrinkage of the ice layers. If scientists do not take those measurements before the ice is gone, "all opportunities will be lost and we will never know what the glacier-ice history in Africa has been", Hardy says. Going higher on the few global sites that still exist, meanwhile, can be dangerous. "It's a risky business if you go to altitudes in this range of 6,000 metres or higher," Schwikowski says

(see 'Climb any mountain').

The ice-core community is discussing ways to save ice for the next generation of scientists, who will have more-advanced theories and measuring tools available to them. In February, Patrick Ginot, a palaeoclimatologist at the Institute of Research for Development (IRD) in Marseilles, France, urged the United Nations Educational, Scientific and Cultural Organization to support a programme that would collect extra ice cores and store them at the Concordia Research Station in central Antarctica. He advocates a "one core for science, two cores for storage" approach that would preserve samples for the future while giving current scientists some to work on.

The IRD has approved a pilot programme in which Ginot will collect three cores from Col du Dôme in the French Alps in 2016, and another three from the mountain of Illimani in Bolivia in 2017. Transporting all that ice to Antarctica will pose logistical challenges, he says, and it will make his work that much more demanding. "With this approach, we have to stay three times longer," he says, to collect three times as many samples as usual.

But persuading science-funding agencies to pay to store samples that will not be used for years could be a hard sell, says Ed Brook, a palaeoclimatologist at the Oregon State University in Corvallis and co-chair of the International Partnerships in Ice Core Sciences, which advocates for ice-core research. Most funding agencies, he says, aim to fund research that is expected to lead quickly to published results. "It's harder," says Brook, "to get funding for longer-term archiving of things." Schwikowski agrees. "I don't know how we could argue, 'We could not publish this year

because we drilled ten ice cores.'"

To encourage science funders to support ice-core storage, the group is working on a report that outlines the importance of preserving records of climate history. Brook expects to have it ready for a major geosciences meeting in 2016. It is important to start the effort soon, he says, because the ice, and the information it contains, is disappearing now. "You're getting rid of the part where we actually have instrumental records to compare and calibrate with," he says. "We don't have that much time."

Still, veteran palaeoclimatologists say that the rapidly changing conditions could prove a boon to the field. Much work needs to be done to understand both the rate of change in ice melt and deposition, and how current climate processes differ from those in the past, when the atmosphere contained much lower concentrations of carbon.

Younger scientists are uncertain how the changes will affect their work. "It definitely makes it harder," says Aron Buffen, a palaeoclimatology doctoral student at Brown University in Providence, Rhode Island, who has worked with Thompson on Quelccaya. If all the ice that formed in years when instruments were measuring weather data disappears, scientists will lose a point of comparison for validating future measurement techniques, he says. A dearth of ice might also discourage custodians of the few remaining samples from sacrificing them to test unproven techniques.

Still, Buffen says that the melting will lead to more questions for research. These include determining which chemical traces will remain behind in sediment and which will return to the atmosphere when the ice melts, as well as distinguishing between melting caused by warmer conditions and sublimation caused by lower humidity. "I wouldn't dissuade anyone from working on tropical glaciers," Buffen says. Future researchers, for instance, could help society to adapt to the changes taking place, if they can provide clues to how shrinking glaciers might affect local ecosystems. And ice at the world's highest spots, as well as in Antarctica and Greenland, will endure for many years to come. Thompson, too, is optimistic about the future, so much so that he offers an online palaeoclimatology course through the Chinese Academy of Sciences. Already, 26 students have enrolled and Thompson hopes that they will go on to study glaciers in the Himalayas.

"It's a bit of a gloomy situation to see these beautiful glaciers going away," says Hardy. "But from the standpoint of careers and science, it presents some interesting opportunities." ∎

Neil Savage *is a freelance writer in Lowell, Massachusetts.*

SARAH IQBAL

TURNING POINT
Arun Shukla

Structural biologist Arun Shukla left his native India for graduate training, as have many other researchers. Unlike most, he worked with three Nobel laureates on two distant continents before returning home. Shukla describes why now is a good time to repatriate to India.

How did you meet your PhD adviser?
While I was in a master's programme in biotechnology at Jawaharlal Nehru University in New Delhi, I was learning about G-protein-coupled receptors (GPCRs), which are involved in almost every physiological process and make up the largest class of potential drug targets. I knew that I wanted to pursue research in this area and attended a fascinating talk by Hartmut Michel, a biochemist at the Max Planck Institute of Biophysics in Frankfurt, Germany, who won the chemistry Nobel in 1988. I spoke with him afterwards and sent him my CV, and he offered me a PhD position.

What was it like at the Max Planck Institute?
It was fun. I was working on expressing GPCRs in different cell types. The goal was to crystallize enough protein to use X-ray diffraction to determine the atomic-level structure, so that we could learn how different drugs bind to these receptors. I realized that this was an area that I could work on for the rest of my life.

Did your PhD work make a mark on the field?
I think so. Crystallizing GPCRs was thought to be impossible at the time. GPCRs are highly mobile proteins that sit in the cell membrane, but for crystallography to be successful you need a stable protein. As a result, their structures were not known. Using nuclear magnetic resonance spectroscopy, we were able to determine the structure of a ligand, a hormone bound to a GPCR. Understanding how a ligand bound to a receptor was a big deal, and the work was published in 2008 as a cover article in *Angewandte Chemie* (J. J. Lopez *et al. Angew. Chem. Int. Edn Engl.* **47,** 1668–1671; 2008). Even today, there are only two such studies in the field. I knew that gaining any insights into GPCR structure would be a landmark and mean a lot to my career.

How did you connect with your next Nobel-laureate mentor?
I was finishing my PhD and knew that I wanted to continue working on GPCRs. Robert Lefkowitz, a biochemist at Duke University in Durham, North Carolina, and future winner of the 2012 chemistry Nobel, is the godfather of GPCRs. I sent him my CV and asked if I could join his lab. Without a formal interview,

he wrote back that I was welcome.

Describe your work in such a competitive field.
The goal — to gain insights into GPCR signalling — was pioneering, and there was a risk of getting scooped. In 2013, Lefkowitz, his Nobel co-recipient Brian Kobilka, and I published the structure of β-arrestin, a GPCR-regulating protein (A. K. Shukla *et al. Nature* **497**, 137–141; 2013). Our paper was in the same issue as one from a group that crystallized a different arrestin.

What prompted you to return to India?
I had watched infrastructure and funding prospects improve in the past decade and thought I could run a better group here given the tight US funding situation, so I started applying for positions. I had several offers, and accepted one at the Indian Institute of Technology in Kanpur.

How is it going?
I have the academic freedom to establish GPCR crystallography as a new line of research in this country, with funding from the Indian Department of Science and Technology and a five-year grant from the Wellcome Trust/Department of Biotechnology India Alliance.

Have there been any roadblocks?
It can take weeks to get reagents and consumables from the United States or Europe. We also lose our top PhD graduates overseas so it can be hard to find a good postdoc. My hope is that if we do good work in India, students will realize that they can stay and have high-impact papers.

What was the best piece of advice you received from the Nobel laureates?
Focus on big questions — do things that are cutting edge and will help to shape the direction of the field. We have to make discoveries, not just publish papers. ∎

INTERVIEW BY VIRGINIA GEWIN

THE BUYOUT

Share and share alike.

BY ANANYO BHATTACHARYA

I knew there was trouble brewing when the envelope was slipped under my cubicle door. I was munching on fortified flakes, at one with my efficiently designed — yet stylish — cubicle (mine in 768 more monthly instalments), when its arrival killed my buzz stone dead. No one sends paper any more unless it's bad news.

I opened it. The summons was no better for being typed on crisp white paper: "Dear Len, Time's up. The board wants to see you pronto. Yours, & c. Ash. B. Mine, Executive Chairman, IOR Life." Not his exact words but I'm smart enough to cut through the commercialese. That's why I've still got a 60% stake in myself. Not like those glazed-eyed 49ers. Everyone knows that once you've crossed that line, your dreams of a buyback are over.

Ping! The call-up was like a flashing red LED inside my head. Boardroom. Forty-five minutes. That was just enough time to jump into my Thermaform shower (my own in just 344 very reasonable weekly payments) and slip into my finest suit; the one with the laser-cut carbon nanorod patches ("for the sharpest elbows in the boardroom"). No way would any of this be mine if I hadn't floated myself on the stock market three years ago. I wasn't born into a top credit rating. Without the IPO, I'd be scraping by at best: no dining out in style, no cubicle and no meds when I got sick. At worst? Out beyond the gates, some man-monster's meat meal. No thank you! Since going public, my credit rating's soared and the board says I might have VP potential.

A short amble down the steel corridors of my complex and I'm outside the boardroom, three minutes early. Right on the second, the bolts slide smoothly over and the ten-inch-thick door swings open.

"Welcome, chief executive!" The chairman's voice booms from the screen at the far end of the room. The other directors beam at me benevolently from their respective positions on the wall. I take the only seat in the room, at the end of the conference table opposite the chairman.

"Now, Len, you've been a *fine* CEO. But lately, we're a little … troubled by the numbers."

"Troubled?" I ventured.

"I'm afraid so. Since we last saw you six months ago, progress has stalled. Shifted units are down on the same quarter last year."

The chairman's voice hardened. "The recovery plan …"

"Which was approved by the board!" I fought to reassert control over my voice, which had risen in pitch by an octave or two.

"… which was approved by the board," he continued sternly, "has not been adequately implemented. There are concerns that you might even become … unproductive."

Unproductive? I thought. *Because of a missed milestone or two?* "That conclusion is completely unwarranted," I said. "Some targets have been missed, yes, but it was the board's decision to raise the price of my units. Times are hard and …"

"Times are hard?" the chairman boomed. Too late, I recognized my misstep. "The economy is booming, Len. What's needed here is more ambition, hard work …"

"A new recovery plan then? With clearly defined milestones …"

"Oh yes, Len. Yes. That, certainly. We want a six-week plan so we can get back on track."

Phew. That would leave little room for manoeuvre but at least …

"And a further buyout of your stock."

I stiffened. "What is the board proposing?"

"Another 9% stake."

I could have wept with joy. That would mean more ops, yes, but it left me with a controlling interest. A 49er? Len? No siree!

The chairman set down the terms of the deal: a generous offer that would allow me to pay off the Thermaform and make 200 more payments on my cubicle. I accepted and the board smiled down on me beatifically.

"Now Len, the nurses are here. We're assured the procedures will be as painless as always." I try to smile. The previous ops had left me with migraines, sudden-onset facial ticks, the occasional blackout …

The nurses roll towards me, their burnished chrome pincers taking a firm grip on my elbows. I'm led away to the ops theatre.

After the op, the board were on hand 24–7 with their guidance, conveniently beamed straight into my head. But the recovery plan had, with hindsight, been far too optimistic.

A scant four weeks later, the white envelope appeared again. This time, the directors were not smiling. The chairman got straight to the point. The board wanted another buyout: 2%. The vote had been unanimous. "You can turn the board down, of course, Len," the chairman said sombrely. "But then we would be forced to dump your stock."

I knew what that meant. My share price would plummet, my credit rating disappear off a cliff. And once the market lost confidence, how long would I keep my job? Goodbye Thermaform. Goodbye cubicle. I'd be man-monster meat for sure.

I nodded dumbly. The nurses were waiting.

Now when the board says "Jump", I jump — well, after the 20 milliseconds it takes for the command to hit the brain stem and travel down the femoral nerve. On the bright side, because the ops were particularly invasive this time, the board gave me two days off. And I do feel happier, although that might have a little to do with the meds they have me on to boost productivity. So what about the name-calling? The kids shouting "zombie!" or "dead eye!" the minute my back's turned? "Rise above it, Len," the chairman says, "and focus on the job." So I do. I barely give it a thought.

And, Lord knows, I certainly don't have too many of *those* any more. ∎

Ananyo Bhattacharya *is a science journalist based in London. He is currently community editor of* The Economist. *Follow him on Twitter: @Ananyo.*
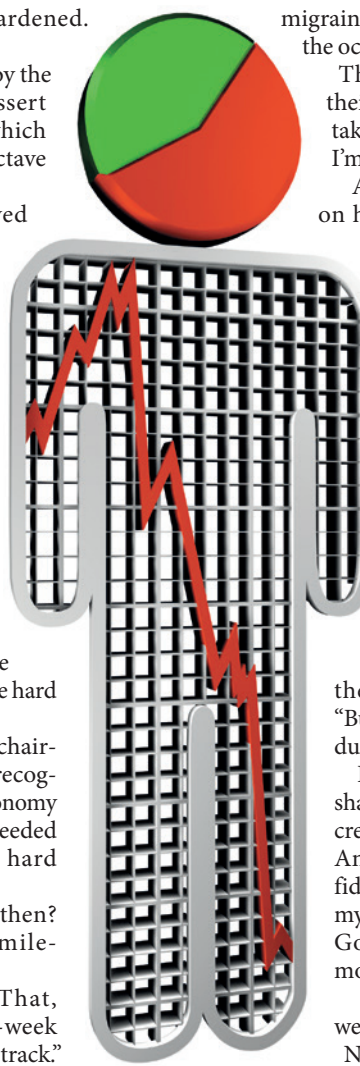
ILLUSTRATION BY JACEY